A Globally Convergent Proximal Newton-Type Method in Nonsmooth Convex Optimization

Boris S. Mordukhovich [*] Xiaoming Yuan[†] Shangzhi Zeng[‡] Jin Zhang[§]

November 16, 2020

Abstract. The paper proposes and justifies a new algorithm of the proximal Newton type to solve a broad class of nonsmooth composite convex optimization problems without strong convexity assumptions. Based on advanced notions and techniques of variational analysis, we establish implementable results on the global convergence of the proposed algorithm as well as its local convergence with superlinear and quadratic rates. For certain structural problems, the obtained local convergence conditions do not require the local Lipschitz continuity of the corresponding Hessian mappings that is a crucial assumption used in the literature to ensure a superlinear convergence of other algorithms of the proximal Newton type. The conducted numerical experiments of solving the l_1 regularized logistic regression model illustrate the possibility of applying the proposed algorithm to deal with practically important problems.

Key words. Nonsmooth convex optimization, machine learning, proximal Newton methods, global and local convergence, metric subregularity

Mathematics Subject Classification (2000) 90C25, 49M15, 49J53

1 Introduction

In this paper we consider a class of optimization problems of the following type:

$$\min_{x\in\mathbb{R}^n} \; F(x) := f(x) + g(x), \tag{1}$$

where both functions $f, g\colon \mathbb{R}^n \to \overline{\mathbb{R}} := (-\infty, \infty]$ are proper, convex, and lower semicontinuous (l.s.c.), while being structurally different from each other. Namely, f is assumed to be twice continuously differentiable with the Lipschitz continuous gradient ∇f on its domain. On the other hand, g is merely continuous on its domain; see Assumption 1.1 for the precise formulations. It has been well recognized that model (1), known as a *composite convex optimization problem*, frequently appears in a variety of applications including, e.g.,

[*] Department of Mathematics, Wayne State University, Detroit, Michigan, USA (boris@math.wayne.edu). Research of this author was partly supported by the USA National Science Foundation under grants DMS-1512846 and DMS-1808978, by the USA Air Force Office of Scientific Research grant #15RT04, and by Australian Research Council under grant DP-190100555.

[†] Department of Mathematics, The University of Hong Kong, Hong Kong, China (xmyuan@hku.hk). This author was supported by the General Research Fund 12302318 from the Hong Kong Research Grants Council.

[‡] Department of Mathematics, The University of Hong Kong, Hong Kong, China (zengsz@connect.hku.hk).

[§] SUSTech International Center for Mathematics, Department of Mathematics, Southern University of Science and Technology, Shenzhen, 518055, China (zhangj9@sustech.edu.cn). Research of this author was supported by National Science Foundation of China 11971220, and by Guangdong Basic and Applied Basic Research Foundation 2019A1515011152.

machine learning, signal processing, and statistics, where f is a *loss function* and g is a *regularizer*; we keep this terminology here. Note that problem (1) contains in fact implicit constraints written as $x \in \Omega := \operatorname{dom} g$.

It is typical in applications that problems of type (1) have a large size, which makes attractive to compute their solutions by employing first-order algorithms such as the *proximal gradient method* (PGM). Given each iteration x^k, the PGM constructs a new iteration x^{k+1} by solving the following optimization subproblem, which approximates the smooth function f in (1) by the linear model:

$$\min_{x \in \mathbb{R}^n} \; l_k(x) := f(x^k) + \nabla f(x^k)^T (x - x^k) + g(x), \tag{2}$$

where ^{T} indicates the matrix transposition. As well known, the PGM applied to (1) generates a sequence of iterates that converges at least sublinearly of rate $O(1/k)$ (see, e.g., [3, 28]) and linearly with respect to the sequence of cost function values—provided that f is strongly convex; see e.g., [34]. Refined results on linear convergence of the PGM are derived under various error bound conditions as in [22, 23, 27, 36, 37].

When f is a twice continuously differentiable function, it is natural to expect algorithms having faster convergence rates by exploiting the Hessian $\nabla^2 f(x^k)$ of f at each iteration x^k and constructing the next iteration x^{k+1} as a solution to the following quadratic subproblem:

$$\min_{x \in \mathbb{R}^n} \; q_k(x) := f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2}(x - x^k)^T H_k (x - x^k) + g(x), \tag{3}$$

where H_k is an appropriate approximation of the Hessian $\nabla^2 f(x^k)$. Methods of this type to solve composite optimization problems (1) are unified under the name of *proximal Newton-type methods*; see, e.g., [20]. To the best of our knowledge, the origin of such methods to solve nonsmooth composite optimization problems given in form (1) can be traced back to the generalized proximal point method developed by Fukushima and Mine [12] who in turn considered it as an extension of Rockafellar's proximal point method [31] to find zeros of maximal monotone operators and subgradient inclusions associated with convex functions. On the other hand, the general scheme of *successive quadratic approximations* to solve optimization-related problems is a common idea of Newton-type and quasi-Newton methods; see the books [10, 16] with their bibliographies. For particular subclasses of composite problems (1), the quadratic approximation scheme (3) contains special versions of the proximal Newton-type methods known as GLMNET [13], newGLMNET [38], QUIC [15], the Newton-LASSO method [29], the projected Newton-type algorithms [34, 35], etc.

Observe further that, due to the convexity of both functions f and g with f being smooth, problem (1) can be equivalently written as the *generalized equation*

$$0 \in \nabla f(x) + \partial g(x) \tag{4}$$

in the sense of Robinson [30], where $\partial g(x)$ is the subdifferential of g at x. Then subproblem (3) for constructing the new iteration x^{k+1} in the proximal Newton method for (4) reduces to solving the following *partially linearized* generalized equation at the given iteration x^k:

$$0 \in \nabla f(x^k) + H_k(x - x^k) + \partial g(x). \tag{5}$$

Various results on the local superlinear and quadratic convergence of iterative sequences $\{x^k\}$ for (5) are obtained in the literature in the framework of quasi-Newton methods for generalized equations under different kinds of regularity conditions imposed on ∂F from (1); see, e.g., the books [7, 10, 16] with the references and discussions therein. In particular, Fischer [11] proposes an iterative procedure to solve generalized equations and proves local superlinear and quadratic convergence of iterates under certain Lipschitz stability property of the corresponding perturbed solution map. More specifically, paper [11] develops a quasi-Newton algorithm to solve (1) in the framework of (5) that exhibits a local superlinear/quadratic convergence in the setting where g is the indicator function of a box constraint, and where H_k in (3) is taken as the regularized Hessian $H_k := \nabla^2 f(x^k) + \alpha_k I$ with $\{\alpha_k\}$ being a positive vanishing sequence satisfying certain conditions. The main assumptions of [11] include the local Lipschitz continuity of the Hessian $\nabla^2 f(x)$ and the upper Lipschitz continuity/calmness of the perturbed solution map (1) at the points in question.

However, how to build a reasonable *globalization* of the local scheme given by (3) has not been completely resolved yet. Various globalizations of the proximal Newton method can be found in the literature, see, e.g., [4, 20, 19, 33]. Unfortunately, all these works require f to be *strongly convex*. In particular, paper by Byrd et al. [4], which addresses the special case of problem (1) with $g := \lambda\|x\|_1$ and $\lambda > 0$, proposes implementable inexactness conditions and backtracking line search procedures to design a globally convergent proximal Newton method, but the local superlinear and quadratic convergence results therein are established under the strong convexity assumption on f. Quite recently [39], the inexactness conditions and backtracking line search procedures of [4] is applied to develop a proximal Newton method for (1) with proving its local convergence of superlinear and quadratic rates by using the Luo-Tseng error bound condition [23] instead of the strong convexity assumption in [4]. However, the convergence results in [39] have a crucial flaw. To achieve a local quadratic convergence rate, the authors of [39] require that parameters of their method satisfy a certain condition involving the constant in the error bound, which is extremely challenging to estimate.

In this paper we design a new globally convergent proximal Newton-type algorithm to solve composite convex optimization problems of class (1) under the following *standing assumptions* on the given data without requiring the strong convexity of the loss function f:

Assumption 1.1. *Impose the following properties of the loss function and the regularizer in* (1)*:*

(i) *Both functions $f, g : \mathbb{R}^n \to (-\infty, \infty]$ are proper, l.s.c., and convex.*

(ii) *The domain of the loss function* $\operatorname{dom} f := \{x \mid f(x) < \infty\}$ *is open, and $f(x)$ is twice continuously differentiable on a closed set* $\Omega \supset (\operatorname{dom} f) \cap (\operatorname{dom} g)$.

(iii) *The regularizer $g(x)$ is continuous on its domain.*

(iv) *The gradient $\nabla f(x)$ is Lipschitz continuous on a closed set Ω from* (i) *with Lipschitz constant $L_1 > 0$.*

(v) *Problem* (1) *has a nonempty solution set denoted by* $\mathcal{X} := \arg\min_{x\in\mathbb{R}^n} F(x)$ *with the optimal value F^*.*

Our main contributions can be summarized as follows:

(1) We develop a *globally convergent* proximal Newton-type algorithm to solve (1) with an *implementable inexact condition* for subproblem (3) and a new reasonable *backtracking line search* strategy. Our line search procedure does not require any restrictive assumptions. It is shown in this way that if the subgradient mapping ∂F is *metrically subregular* at any limiting point of the iterative sequence, the backtracking line search procedure accepts a unit step size when the iterates are closed to the solution. Furthermore, we prove that the proposed proximal Newton-type algorithm exhibits a *local convergence* with the *quadratic convergence rate*. Numerical experiments are performed to solve the l_1 *regularized logistic regression* problem that illustrate the efficiency of the proposed algorithm.

(2) We establish novel local convergence results for the proposed algorithm under the *metric q-subregularity* assumption imposed on the subgradient mapping ∂F for any positive number q. If $q \in (0,1)$, the obtained results require *less restrictive assumptions* in comparison with the case of metric subregularity ($q = 1$) to ensure a superlinear convergence of iterates, while for $q > 1$ we achieve a convergence rate that is *higher than quadratic.*

(3) When the loss function f in (1) satisfies additional structural assumptions, we obtain a local superlinear convergence rate of our proposed algorithm *without imposing the Lipschitz continuity* of the Hessian matrix $\nabla^2 f(x)$. The latter assumption is crucial for establishing a fast convergence of the previously known algorithms of the proximal Newton type.

The rest of the paper is organized as follows. Section 2 briefly overviews the notions and results of variational analysis needed for the subsequent material. In Section 3 we present our proximal Newton-type algorithm and establish its global convergence. Section 4 contains results on the local superlinear and quadratic convergence of the proposed algorithm under the metric subregularity assumption on the subgradient mapping. In Section 5 we derive advanced local convergence results under the metric q-subregularity conditions imposed on ∂F considering separately the cases where $q \in (0,1)$ and where $q > 1$. The next Section 6 is devoted to problem (1) with a certain structure of the loss function f and establishes in this case a superlinear convergence of the proposed algorithm without the Lipschitz continuity of the loss function Hessian. Finally, Section 7 conducts and analyzes numerical experiments to solve the practically important l_1 regularized logistic regression problem by implementing the designed proximal Newton-type method.

2 Preliminaries from Variational Analysis

Here we recall and discuss some material from variational analysis that is broadly used in what follows. The reader can find more details and references in the books [7, 24, 32].

Throughout the paper we use the standard notation. Recall that $\mathbb{R}^n$ signifies an n-dimensional Euclidean space with the inner product $\langle \cdot, \cdot \rangle$ and the norm denoted by $\| \cdot \|$, while the 1-norm is signified by $\| \cdot \|_1$. For any matrix $A \in \mathbb{R}^{m \times n}$ we have $\|A\| := \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$ with $\tilde{\sigma}_{\min}(A)$ standing for the smallest nonzero singular value of A. The symbols $\mathbb{B}_r(x)$ and $\overline{\mathbb{B}}_r(x)$ denote the open and the closed Euclidean norm ball centered at x with radius $r > 0$, respectively, while we use $\mathbb{B}$ and $\overline{\mathbb{B}}$ for the corresponding unit balls around the origin. Given a nonempty subset $\Omega \subset \mathbb{R}^n$, denote by $\operatorname{bd} \Omega$ its boundary and consider the associated distance function $\operatorname{dist}(x; \Omega) := \inf\{\|x - y\| \,\big|\, y \in \Omega\}$ and the indicator function $\delta_\Omega(x)$ equal 0 if $x \in \Omega$ and ∞ otherwise. The graph of a set-valued mapping/multifunction $\Psi \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is given by $\operatorname{gph} \Psi := \{(x, v) \in \mathbb{R}^n \times \mathbb{R}^m \mid v \in \Psi(x)\}$, and the inverse to Ψ is $\Psi^{-1}(v) := \{x \in \mathbb{R}^n \mid v \in \Psi(x)\}$.

The following fundamental properties of set-valued mappings are employed in the paper to establish fast local convergence results for the proposed proximal Newton-type algorithm.

Definition 2.1. *Let $\Psi \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ be a set-valued mapping, let $(\bar{x}, \bar{v}) \in \operatorname{gph} \Psi$, and let $q > 0$.*

(i) *We say that Ψ is* METRICALLY q-SUBREGULAR *at $(\bar{x}, \bar{v})$ with modulus $\kappa > 0$ if there is $\varepsilon > 0$ such that*

$$\operatorname{dist}\big(x; \Psi^{-1}(\bar{v})\big) \leq \kappa \operatorname{dist}\big(\bar{v}; \Psi(x)\big)^q \quad \text{for all } \; x \in \mathbb{B}_\varepsilon(\bar{x}). \tag{6}$$

(ii) *Ψ is said to be* METRICALLY SUBREGULAR *at $(\bar{x}, \bar{v})$ if $q = 1$ in (6).*

The metric subregularity property has been well recognized and applied in variational analysis and optimization numerical aspects. The reader can find more information and references in [7, 24] with the commentaries and the bibliographies therein. In this paper we employ metric subregularity of *subgradient mappings*, which form a remarkable class of multifunctions with special properties. Various sufficient conditions and characterizations of this property of subgradient mappings are given in [1, 2, 9] in terms of certain second-order growth conditions imposed on the function in question.

The metric q-subregularity of order $q \in (0,1)$, known also as *Hölder metric subregularity*, is much less investigated, while some verifiable conditions for the fulfillment of this property can be found in, e.g., [14, 21, 40]. Note that the Hölder metric subregularity is clearly a weaker assumption in comparison with the standard metric subregularity property.

The case of *higher-order metric subregularity* with $q > 1$ in (6) is largely open in the literature. One of the reasons for this is that the corresponding *metric q-regularity* property with $q > 1$ does not make sense, since it holds only for constant mappings. Nevertheless, it is shown in [25] that the higher-order metric subregularity is a useful property in variational analysis and optimization. This property is characterized for subgradient mappings in [25] via a higher-order growth condition, and its *strong* version is applied therein to the convergence analysis of quasi-Newton methods for generalized equations.

Next we consider the *proximal mapping*

$$\mathrm{Prox}_g(u) := \mathrm{argmin}\Big\{ g(x) + \frac{1}{2}\|x-u\|^2 \ \Big|\ x \in \mathbb{R}^n \Big\}, \quad u \in \mathbb{R}^n, \tag{7}$$

associated with a proper function $g\colon \mathbb{R}^n \to \overline{\mathbb{R}}$. A crucial role of proximal mappings has been well recognized not only in proximal Newton-type algorithms (see, e.g., [4, 20]), but also in other second-order methods of numerical optimization. In particular, we refer the reader to the very recent papers [17, 26], where the proximal mappings are used for designing superlinearly convergent Newton-type algorithms to find tilt-stable local minimizers of nonconvex extended-real-valued functions and to solve subgradient inclusions in a large generality. If g is l.s.c. and convex, then the proximal mapping (7) is single-valued and nonexpansive on $\mathbb{R}^n$, i.e., Lipschitz continuous with constant one; see, e.g., [32, Theorem 12.12].

It is important to emphasize that in many practical models of type (1) arising, in particular, in machine learning and statistics, the proximal mapping associated with the regularizer term g (e.g., when g is the l_1-norm, the group Lasso regularizer, etc.) can be easily computed. This is the case of the l_1 regularized logistic regression problem in our applications developed in Section 7.

Having (7), define further the *prox-gradient mapping* associated with (1) by

$$r(x) := x - \mathrm{Prox}_g\big(x - \nabla f(x)\big), \quad x \in \mathbb{R}^n, \tag{8}$$

and present some properties of (8) used in what follows. The first proposition is a combination of Theorem 3.4 and Theorem 3.5 established in [8].

Proposition 2.1. *Let ∇f be Lipschitz continuous with modulus L_1 around $\bar{x}$, and let the mapping $\nabla f(x) + \partial g(x)$ be metrically subregular at $(\bar{x}, 0)$, i.e., there exist numbers $\varepsilon, \kappa > 0$ such that*

$$\mathrm{dist}(x; \mathcal{X}) \le \kappa\, \mathrm{dist}\big(0; \nabla f(x) + \partial g(x)\big) \ \textit{ for all } \ x \in \mathbb{B}_\varepsilon(\bar{x}).$$

Then whenever $x \in \mathbb{B}_\varepsilon(\bar{x})$ we have the estimate

$$\mathrm{dist}(x; \mathcal{X}) \le (1+\kappa)(1+L_1)\|r(x)\|.$$

The next proposition give a reverse statement to Proposition 2.1 while providing an estimate of the norm of (8) via the distance to the solution set of the convex composite problem (1).

Proposition 2.2. *Let ∇f be Lipschitz continuous with modulus L_1 on $\mathbb{R}^n$. Then we have the estimate*

$$\|r(x)\| \le (2+L_1)\mathrm{dist}(x;\mathcal{X}) \ \ \textit{for all} \ \ x \in \mathbb{R}^n.$$

Proof. Observe first that the mapping $r(x)$ is well-defined and single-valued for all $x \in \mathbb{R}^n$ due to the aforementioned result of [32]. It easily follows from Assumption 1.1 that the nonempty solution set $\mathcal{X}$ is closed and convex; hence each point $x \in \mathbb{R}^n$ has the unique projection $\pi_x \in \mathcal{X}$ onto $\mathcal{X}$. Note that $\pi_x - \mathrm{Prox}_g(\pi_x - \nabla f(\pi_x)) = 0$ for $\pi_x \in \mathcal{X}$. Thus we verify the claim of the proposition by

$$\begin{aligned}\|r(x)\| &= \big\|x - \mathrm{Prox}_g\big(x - \nabla f(x)\big)\big\| \\ &= \big\|x - \mathrm{Prox}_g\big(x - \nabla f(x)\big) - \big(\pi_x - \mathrm{Prox}_g(\pi_x - \nabla f(\pi_x))\big)\big\| \\ &\le \|x - \pi_x\| + \big\|\mathrm{Prox}_g\big(x - \nabla f(x)\big) - \mathrm{Prox}_g\big(\pi_x - \nabla f(\pi_x)\big)\big\| \\ &\le \|x - \pi_x\| + \big\|\big(x - \nabla f(x)\big) - (\pi_x - \nabla f(\pi_x))\big\| \\ &\le (2+L_1)\|x - \pi_x\|, \quad x \in \mathbb{R}^n,\end{aligned}$$

where the second inequality holds since the proximal mapping Prox_g is nonexpansive. □

Finally, we obtain an extension of Proposition 2.1 to case where the subgradient mapping $\nabla f + \partial g$ in (1) satisfies the Hölder subregularity property in the point in question.

Proposition 2.3. *Let ∇f be Lipschitz continuous with modulus L_1 around $\bar{x}$, and let the mapping $\nabla f(x) + \partial g(x)$ be metrically q-subregular at $(\bar{x}, 0)$ with $q \in (0,1]$, i.e., there exist $\varepsilon_1, \kappa_1 > 0$ such that*

$$\mathrm{dist}(x;\mathcal{X}) \le \kappa_1 \mathrm{dist}\big(0; \nabla f(x) + \partial g(x)\big)^q \ \ \textit{for all} \ \ x \in \mathbb{B}_{\varepsilon_1}(\bar{x}).$$

Then we find constants $\varepsilon_2, \kappa_2 > 0$ that ensure the estimate

$$\mathrm{dist}(x;\mathcal{X}) \le \kappa_2 \|r(x)\|^q \ \ \textit{whenever} \ \ x \in \mathbb{B}_{\varepsilon_2}(\bar{x}). \tag{9}$$

Proof. By (8) we have the inclusions

$$r(x) \in \nabla f(x) + \partial g\big(x - r(x)\big) \ \ \text{and}$$

$$r(x) + \nabla f\big(x - r(x)\big) - \nabla f(x) \in \nabla f\big(x - r(x)\big) + \partial g\big(x - r(x)\big)$$

for all $x \in \mathbb{R}^n$. When $x \in \mathbb{B}_\varepsilon(\bar{x})$, it follows from the imposed assumption that

$$\begin{aligned}\mathrm{dist}\big(x - r(x);\mathcal{X}\big) &\le \kappa \, \mathrm{dist}\big(0; r(x) + \nabla f\big(x - r(x)\big) - \nabla f(x)\big)^q \\ &\le \kappa(1+L_1)^q \|r(x)\|^q,\end{aligned}$$

which lead us to the resulting estimates for such x:

$$\begin{aligned}\mathrm{dist}(x;\mathcal{X}) &\le \mathrm{dist}\big(x - r(x);\mathcal{X}\big) + \|r(x)\| \\ &\le 1 + \kappa(1+L_1)^q) \max\big\{\|r(x)\|, \|r(x)\|^q\big\}.\end{aligned}$$

Applying now Proposition 2.2 tells us that, whenever $\text{dist}(x;\mathcal{X}) \le 1/(2+L_1)$ and $x \in \mathbb{R}^n$, we get

$$\|r(x)\| \le (2+L_1)\text{dist}(x;\mathcal{X}) \le 1.$$

Letting $\varepsilon_2 := \min\{1/(2+L_1), \varepsilon_1\}$ and remembering that $q \le 1$ bring us to the inequality

$$\text{dist}(x;\mathcal{X}) \le (1+\kappa(1+L_1)^q)\|r(x)\|^q \text{ for all } x \in \mathbb{B}_{\varepsilon_2}(\bar{x}),$$

which verifies (9) with $\kappa_2 := (1+\kappa(1+L_1)^q)$ and thus completes the proof of the proposition. □

3 The New Algorithm and Its Global Convergence

In this section we describe the proposed proximal Newton-type algorithm to solve the class of composite convex optimization problems (1) with justifying its global convergence under the standing assumptions.

Given a current iteration x^k for each $k = 0, 1, \ldots$, we select a *positive semidefinite matrix* B_k as an arbitrary approximation of the Hessian $\nabla^2 f(x^k)$ satisfying the *standing boundedness assumption*:

$$\text{there exists } M \ge 0 \text{ such that } \|B_k\| \le M \text{ whenever } k = 0, 1, \ldots. \tag{10}$$

If the gradient mapping ∇f is uniformly Lipschitz continuous along the sequence of iterations with constant L_1, then (10) holds for $B_k = \nabla^2 f(x^k)$ with $M = L_1$. In the general case of B_k, pick any constants $c > 0$ and $\rho \in (0,1]$ and, using the prox-regular mapping (8), consider the positive number $\alpha_k := c\|r(x^k)\|^\rho$ and define the *quasi-Newton approximation* of the Hessian of f at x^k by

$$H_k := B_k + \alpha_k I \text{ for all } k = 0, 1, \ldots, \tag{11}$$

which is a positive definite matrix. Then similarly to [4], but with the different approximation (11), denote

$$r_k(x) := x - \text{Prox}_g\left(x - \nabla f(x^k) - H_k(x - x^k)\right) \tag{12}$$

and select $\hat{x}^k$ as an *approximate minimizer* of the quadratic *subproblem* for (1) given by

$$\min_{x \in \mathbb{R}^n} q_k(x) := f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{2}(x - x^k)H_k(x - x^k) + g(x) \tag{13}$$

with the *residual* number $\|r_k(\hat{x}^k)\|$ measuring the approximate optimality of $\hat{x}^k$ in (13). Observing that $\|r_k(\hat{x}^k)\| = 0$ if and only if $\hat{x}^k$ is an *exact* solution to subproblem (13), we use the nonnegative number

$$\hat{r}_k := \min\left\{\|r_k(\hat{x}^k)\|, \text{dist}\big(0; \partial q_k(\hat{x}^k)\big)\right\}$$

with $r_k(x)$ taken from (12) as the *optimality measure* of $\hat{x}^k$ in subproblem (13). Adapting the scheme of [4] in our new setting, let us impose the following two estimates as *inexact conditions* for choosing $\hat{x}^k$ as an approximate solution to subproblem (13):

$$\min\left\{\|r_k(\hat{x}^k)\|, \text{dist}\big(0; \partial q_k(\hat{x}^k)\big)\right\} \le \eta_k \|r(x^k)\| \text{ and } q_k(\hat{x}^k) \le q_k(x^k) \tag{14}$$

with the parameter $\eta_k := \nu \min\{1, \|r(x^k)\|^\varrho\}$ defined via (12) and some numbers $\nu \in [0, 1/2)$ and $\varrho > 0$.

Using the above constructions and the line search procedure inspired by [5, 6], we are ready to propose the proximal Newton-type algorithm designed as follows:

Algorithm 1 Proximal Newton-type method

```
Choose $x^0 \in \mathbb{R}^n$, $0 < \theta, \sigma, \gamma < 1$, $C > F(x^0)$, $c > 0$, and $\rho \in (0, 1]$.
for $k = 0, 1, \ldots$ do
    1. Update the approximation of the Hessian matrix $B_k$.
    2. Form the quadratic model (3) with $H_k := B_k + \alpha_k I$ and $\alpha_k := c\|r(x^k)\|^\rho$.
    3. Obtain an inexact solution $\hat{x}^k$ of (3) satisfying the conditions in (14).
    4. If $k = 0$, let $\vartheta_1 := r(x^0)$ and go to Step 5. For $k \geq 1$, if $\|r(\hat{x}^k)\| \leq \sigma\vartheta_k$ and $f(\hat{x}^k) \leq C$, let $t_k := 1$,
       $\vartheta_{k+1} := \|r(\hat{x}^k)\|$, and go to Step 6. Otherwise, let $\vartheta_{k+1} := \vartheta_k$ and go to Step 5.
    5. Perform a backtracking line search along the direction $d^k := \hat{x}^k - x^k$ by setting $t_k := \gamma^{m_k}$, where
       $m_k$ is the smallest nonnegative integer $m$ such that
           $$F(x^k + \gamma^m d^k) \leq F(x^k) - \theta\alpha_k\gamma^m\|d^k\|^2. \qquad (15)$$
    6. Set $x^{k+1} := x^k + t_k d^k$.
end for
```

In the rest of this section we show that the proposed algorithm *globally converges* under the mild standing assumptions, which are imposed above and will not be repeated. Let us start with the following lemma providing a subgradient estimate for subproblem (13) at the approximate solution.

Lemma 3.1. *Given an approximate solution $\hat{x}^k$ to* (13)*, there exists a vector $e_k \in \mathbb{R}^n$ such that*

$$e_k \in \nabla f(x^k) + H_k(\hat{x}^k - x^k) + \partial g(\hat{x}^k - e_k) \quad \textit{and} \quad \|e_k\| \leq \nu \min\left\{\|r(x^k)\|, \|r(x^k)\|^{1+\varrho}\right\}. \tag{16}$$

Proof. Let $e_k := r_k(\hat{x}^k) = \hat{x}^k - \operatorname{Prox}_g(\hat{x}^k - \nabla f(x^k) - H_k(\hat{x}^k - x^k))$ and pick any $\zeta_k \in \partial q_k(\hat{x}^k)$. Then we have

$$e_k \in \nabla f(x^k) + H_k(\hat{x}^k - x^k) + \partial g(\hat{x}^k - e_k) \quad \text{and} \quad \zeta_k \in \nabla f(x^k) + H_k(\hat{x}^k - x^k) + \partial g(\hat{x}^k),$$

which follow from (7) and the subdifferential sum rule of convex analysis. Since the subgradient mapping ∂g is monotone, this readily tells us that

$$\langle e_k - \zeta_k, -e_k \rangle \geq 0,$$

which in turn yields the estimates

$$\|e_k\|^2 \leq \langle \zeta_k, e_k \rangle \leq \|e_k\| \cdot \|\zeta_k\|.$$

Using finally the inexact conditions (14) for $\hat{x}^k$, we arrive at the claim of the lemma. □

The next lemma provides elaborations on Step 5 of the proposed algorithm with the decreasing of the cost function in (1) by the backtracking line search.

Lemma 3.2. *Let t_k be chosen by the backtracking line search in Step* 5 *of Algorithm* 1 *at iteration k. Then we have the size estimate*

$$t_k \geq \frac{\gamma(1-\theta)\alpha_k}{L_1} \tag{17}$$

with the cost function decrease satisfying

$$F(x^{k+1}) - F(x^k) \leq -\frac{\gamma\theta(1-\theta)}{2L_1}\left(\frac{(1-2\nu)\alpha_k}{(1+\nu)(1+M+\alpha_k)}\right)^2 \|r(x^k)\|^2. \tag{18}$$

Proof. Since $\hat{x}^k$ is an inexact solution to (3) obeying the conditions in (14), it follows that

$$0 \geq q_k(\hat{x}^k) - q_k(x^k) = l_k(\hat{x}^k) - l_k(x^k) + \frac{1}{2}(\hat{x}^k - x^k)^T H_k(\hat{x}^k - x^k),$$

where l_k is the linear part of q_k defined in (2). This yields

$$l_k(x^k) - l_k(\hat{x}^k) \geq \frac{1}{2}(\hat{x}^k - x^k)^T H_k(\hat{x}^k - x^k) \geq \frac{1}{2}\alpha_k\|\hat{x}^k - x^k\|^2. \tag{19}$$

By $r(x^k) = x^k - \text{Prox}_g(x^k - \nabla f(x^k))$ we deduce from the stationary and subdifferential sum rules that

$$r(x^k) \in \nabla f(x^k) + \partial g\big(x^k - r(x^k)\big).$$

Furthermore, Lemma 3.1 gives us the condition $e_k \in \nabla f(x^k) + H_k(\hat{x}^k - x^k) + \partial g(\hat{x}^k - e_k)$ for $\hat{x}^k$ with e_k satisfying the estimate $\|e_k\| \leq \nu\|r(x^k)\|$. The monotonicity of the subgradient mapping ∂g ensures that

$$\big\langle r(x^k) + H_k(\hat{x}^k - x^k) - e_k, x^k - r(x^k) - \hat{x}^k + e_k\big\rangle \geq 0,$$

which therefore leads us to the inequality

$$\|r(x^k)\|^2 + \|e_k\|^2 + (\hat{x}^k - x^k)^T H_k(\hat{x}^k - x^k) \leq \big\langle r(x^k), x^k - \hat{x}^k + H_k(x^k - \hat{x}^k)\big\rangle + \big\langle e_k, 2r(x^k) + (\hat{x}^k - x^k) + H_k(\hat{x}^k - x^k)\big\rangle.$$

Using again the condition $\|e_k\| \leq \nu\|r(x^k)\|$ together with $\|B_k\| \leq M$ from (10) results in

$$\|r(x^k)\|^2 \leq (1+\nu)(1+M+\alpha_k)\|r(x^k)\| \cdot \|\hat{x}^k - x^k\| + 2\nu\|r(x^k)\|^2.$$

Remembering the choice of $\nu \in [0, \frac{1}{2})$, we estimate the prox-gradient mapping (8) at the iteration x^k by

$$\|r(x^k)\| \leq \frac{(1+\nu)(1+M+\alpha_k)}{1-2\nu}\|\hat{x}^k - x^k\|. \tag{20}$$

Next let us show that the backtracking line search along the direction $d^k = \hat{x}^k - x^k$ in Step 5 is well-defined and the proposed step size ensures a sufficient decrease in the cost function F. It follows from the Lipschitz continuity of ∇f that

$$f(x^k + \tau d^k) \leq f(x^k) + \tau\nabla f(x^k)^T d^k + \frac{L_1}{2}\tau^2\|d^k\|^2 \quad \text{for any} \quad \tau \geq 0,$$

and thus we deduce from the definition of l_k in (2) that

$$F(x^k) - F(x^k + \tau d^k) \geq l_k(x^k) - l_k(x^k + \tau d^k) - \frac{L_1}{2}\tau^2\|d^k\|^2.$$

This implies by the convexity of g that

$$l_k(x^k) - l_k(x^k + \tau d^k) \geq \tau\big(l_k(x^k) - l_k(x^k + d^k)\big).$$

Combining the latter with (19) and using the choice of $\theta \in (0,1)$ yield the relationships

$$\begin{aligned}
& F(x^k) - F(x^k + \tau d^k) - \frac{\theta\alpha_k\tau}{2}\|d^k\|^2 \\
\geq\ & l_k(x^k) - l_k(x^k + \tau d^k) - \frac{L_1}{2}\tau^2\|d^k\|^2 - \frac{\theta\alpha_k\tau}{2}\|d^k\|^2 \\
\geq\ & \tau\big(l_k(x^k) - l_k(x^k + d^k)\big) - \frac{L_1}{2}\tau^2\|d^k\|^2 - \frac{\theta\alpha_k\tau}{2}\|d^k\|^2 \\
\geq\ & (1-\theta)\tau\frac{\alpha_k}{2}\|d^k\|^2 - \frac{L_1}{2}\tau^2\|d^k\|^2 \\
=\ & \frac{\tau}{2}\|d^k\|^2\big((1-\theta)\alpha_k - L_1\tau\big).
\end{aligned} \tag{21}$$

This tells us that the backtracking line search criterion (15) fulfills when $0 < \tau \leq \frac{(1-\theta)\alpha_k}{L_1}$, and thus the step size t_k satisfies the claimed condition (17). Substituting now $\tau := t_k \geq \frac{\gamma(1-\theta)\alpha_k}{L_1}$ into (21) and employing the estimate of $\|r(x^k)\|^2$ from (20), we arrive at the inequalities

$$\begin{aligned} F(x^k) - F(x^k + t_k d^k) &\geq \frac{\theta \alpha_k t_k}{2}\|d^k\|^2 \\ &\geq \frac{\gamma\theta(1-\theta)\alpha_k^2}{2L_1}\left(\frac{(1-2\nu)}{(1+\nu)(1+M+\alpha_k)}\right)^2 \|r(x^k)\|^2, \end{aligned}$$

which verify the decreasing condition (18) and thus completes the proof of the lemma. □

Now we are ready to prove the global convergence of Algorithm 1. Define the sets

$$K := \{0, 1, \dots\} \ \text{ and } \ K_0 := \{0\} \cup \{k+1 \in K \mid \text{Step 5 is not applied at iteration } k\}. \tag{22}$$

Theorem 3.1. *Let $\{x^k\}$ be the sequence of iterates generated by Algorithm 1 with an arbitrarily chosen stating point $x^0 \in \mathbb{R}^n$ under the standing assumptions made, and let the set K_0 is defined in (22). Then K_0 is infinite and we have the residual condition*

$$\liminf_{k\to\infty} \|r(x^k)\| = 0 \tag{23}$$

along the prox-gradient mapping (8). Furthermore, the boundedness of $\{x^k\}$ yields the convergence to the optimal value $\lim_{k\to\infty} F(x^k) = F^*$ and ensures that any limiting point of $\{x^k\}$ is a global minimizer in (1).

Proof. First we verify that the set K_0 is infinite. Arguing by contraposition, suppose that K_0 is finite and denote $\bar{k} := \max_{k\in K_0} k$. It follows from Lemma 3.2 that for any $k > \bar{k}$ we get

$$F(x^{k+1}) - F(x^k) \leq -\frac{\gamma\theta(1-\theta)}{2L_1}\left(\frac{(1-2\nu)\alpha_k}{(1+\nu)(1+M+\alpha_k)}\right)^2 \|r(x^k)\|^2,$$

which tells us therefore that

$$\sum_{k=\bar{k}}^{\infty} \frac{\gamma\theta(1-\theta)}{2L_1}\left(\frac{(1-2\nu)\alpha_k}{(1+\nu)(1+M+\alpha_k)}\right)^2 \|r(x^k)\|^2 \leq F(x^{\bar{k}}) - F^* \leq 0.$$

The latter implies in turn that

$$\lim_{k\to\infty} \frac{\gamma\theta(1-\theta)}{2L_1}\left(\frac{(1-2\nu)\alpha_k}{(1+\nu)(1+M+\alpha_k)}\right)^2 \|r(x^k)\|^2 = 0.$$

Remembering the choice of $\alpha_k = c\|r(x^k)\|^\rho$ with $c, \rho > 0$ ensures that

$$\lim_{k\to\infty} \|r(x^k)\| = 0,$$

and hence there exists $k > \bar{k}$ such that $k \in K_0$; a contradiction showing that the set K_0 is infinite.

Thus we can reorganize K_0 in such a way that $0 = k_0 < k_1 < k_2 < \dots$. It follows from Step 4 of Algorithm 1 that the estimate

$$\|r(x^{k_{\ell+1}})\| \leq \sigma \|r(x^{k_\ell})\| \ \text{ whenever } \ \ell = 0, 1, \dots$$

holds with the chosen number $\sigma \in (0,1)$ in the algorithm, and therefore we get

$$0 \le \limsup_{\ell\to\infty} \|r(x^{k_\ell})\| \le \lim_{\ell\to\infty} \sigma^\ell \|r(x^{k_0})\| = 0,$$

which clearly yields (23). The continuity of $r(\cdot)$ ensures that $\|r(\bar{x})\| = 0$ for a limiting point $\bar{x}$ of the sequence $\{x^k\}_{k\in K_0}$, and thus $\bar{x} \in \mathcal{X}$. Consider now any limiting point $\bar{x}$ of the entire sequence of iterates $\{x^k\}_{k\in K}$. If there exists $\bar{k}$ such that $k \in K_0$ for all $k \ge \bar{k}$, it is easy to see that $\bar{x}$ is a global minimizer of (1). Otherwise, for any $k \notin K_0$ denote by $k_\ell \in K_0$ the largest number satisfying $k_\ell < k$ and hence get

$$F^* \le F(x^k) \le F(x^{k-1}) \le \ldots \le F(x^{k_\ell}).$$

Since the sequence $\{x^k\}_{k\in K_0}$ is bounded, and since any limiting point of $\{x^k\}_{k\in K_0}$ is a global minimizer in (1) as already shown, it follows that $\lim\limits_{\ell\to\infty} F(x^{k_\ell}) = F^*$. This readily verifies by the constructions above that $\lim\limits_{k\to\infty} F(x^k) = F^*$, and thus any limiting point of $\{x^k\}_{k\in K}$ provides a global minimum to (1). □

We conclude this section with a consequence of Theorem 3.1 giving an easily verifiable condition for the boundedness of the sequence of iterates in Algorithm 1. Recall that a function $\varphi\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is *coercive* if $\varphi(x) \to \infty$ provided that $\|x\| \to \infty$.

Corollary 3.2. *In addition to the standing assumptions imposed above, suppose that the cost function F in* (1) *is coercive. Then we have $\lim\limits_{k\to\infty} F(x^k) = F^*$ for the sequence of iterates $\{x^k\}$ generated by Algorithm* 1*, and any limiting point of $\{x^k\}$ is a global minimizer in* (1)*.*

Proof. According to Steps 4 and 5 of Algorithm 1, the sequence $\{x^k\}$ generated by the algorithm satisfies the condition $F(x^k) \le C$ for all k. Then the coercivity of F implies that the sequence $\{x^k\}$ is bounded. Thus we deduce the conclusions of the corollary from Theorem 3.1. □

4 Quadratic Local Convergence under Metric Subregularity

This section provides a detailed study of the local convergence of Algorithm 1 under the metric regularity assumption on the subgradient mapping in (1); see Section 2 for the discussion of this property. The main result here establishes superlinear local convergence rates depending on the selected exponent $\rho \in (0,1]$ in the algorithm, which gives us the quadratic convergence in the case where $\rho = 1$. Our analysis partly follows the scheme of [11] for a Newtonian algorithm to solve generalized equations with nonisolated solutions under certain Lipschitzian properties of perturbed solution sets. Note that the imposed metric subregularity allows us to avoid limitations of the line search procedure (needed for establishing the global convergence of our algorithm in Section 2 that is not addressed in [11]) to achieve now the fast local convergence.

Let us start with the following lemma giving us a norm estimate of directions d^k in Algorithm 1.

Lemma 4.1. *Let $\{x^k\}$ be the sequence generated by Algorithm* 1 *under the standing assumptions, and let $\bar{x} \in \mathcal{X}$ be any limiting point of $\{x^k\}$. If the Hessian $\nabla^2 f$ is locally Lipschitzian around $\bar{x}$, then there exist numbers $\varepsilon, L_2 > 0$ such that we have the estimate*

$$\|d^k\| \le \frac{1}{\alpha_k}\Big(\frac{L_2}{2}\mathrm{dist}(x^k;\mathcal{X})^2 + \|B_k - \nabla^2 f(x^k)\|\,\mathrm{dist}(x^k;\mathcal{X}) + 2\alpha_k \mathrm{dist}(x^k;\mathcal{X}) + (1+M+\alpha_k)\nu\|r(x^k)\|^{1+\varrho}\Big)$$

for the directions $d^k := \hat{x}^k - x^k$, provided that $x^k \in \mathbb{B}_\varepsilon(\bar{x})$.

Proof. Remembering that $\hat{x}^k$ is an inexact solution to (3) satisfying conditions (14), we apply Lemma 3.1 and find a vector e^k such that the relationships in (16) hold. Denoting by π_x^k the (unique) projection of x^k onto the solution map $\mathcal{X}$, we get by basic convex analysis that $0 \in \nabla f(\pi_x^k) + \partial g(\pi_x^k)$ and thus

$$\nabla f(x^k) - \nabla f(\pi_x^k) + H_k(\pi_x^k - x^k) \in \nabla f(x^k) + H_k(\pi_x^k - x^k) + \partial g(\pi_x^k).$$

Since the mapping $\nabla f(x^k) + H_k(\cdot - x^k) + \partial g(\cdot)$ is strongly monotone on $\mathbb{R}^n$ with constant α_k, we have

$$\langle \nabla f(x^k) - \nabla f(\pi_x^k) + H_k(\pi_x^k - x^k) - e_k + H_k e_k, \pi_x^k - \hat{x}^k + e_k\rangle \geq \alpha_k \|\pi_x^k - \hat{x}^k + e_k\|^2. \tag{24}$$

Combining the above with the algorithm constructions gives us the estimates

$$\begin{aligned}
\|\pi_x^k - \hat{x}^k + e_k\| &\leq \frac{1}{\alpha_k}\|\nabla f(x^k) - \nabla f(\pi_x^k) + H_k(\pi_x^k - x^k) - e_k + H_k e_k\| \\
&\leq \frac{1}{\alpha_k}\big(\|\nabla f(x^k) + \nabla^2 f(x^k)(\pi_x^k - x^k) - \nabla f(\pi_x^k)\| + \|(H_k - \nabla^2 f(x^k))(\pi_x^k - x^k)\| + \|e_k - H_k e_k\|\big) \\
&\leq \frac{1}{\alpha_k}\Big(\|\nabla f(x^k) + \nabla^2 f(x^k)(\pi_x^k - x^k) - \nabla f(\pi_x^k)\| + \|B_k - \nabla^2 f(x^k)\| \cdot \|x^k - \pi_x^k\| \\
&\quad + \alpha_k \|x^k - \pi_x^k\| + (1+M)\|e_k\|\Big) \\
&\leq \frac{1}{\alpha_k}\Big(\|\nabla f(x^k) + \nabla^2 f(x^k)(\pi_x^k - x^k) - \nabla f(\pi_x^k)\| + \|B_k - \nabla^2 f(x^k)\| \operatorname{dist}(x^k; \mathcal{X}) \\
&\quad + \alpha_k \operatorname{dist}(x^k; \mathcal{X}) + (1+M)\nu\|r(x^k)\|^{1+\varrho}\Big),
\end{aligned}$$

where the third inequality follows from the choice of $H_k = B_k + \alpha_k I$ while the fourth inequality is implied by $\|e_k\| \leq \nu\|r(x^k)\|^{1+\varrho}$. Since the Hessian mapping $\nabla^2 f$ is locally Lipschitzian around $\bar{x}$, there exist positive numbers ε and L_2 such that for any $x, y \in \mathbb{B}_\varepsilon(\bar{x})$ we get

$$\|\nabla f(x) + \nabla^2 f(x)(y - x) - \nabla f(y)\| \leq \frac{L_2}{2}\|x - y\|^2.$$

Furthermore, it follows from $x^k \in \mathbb{B}_\varepsilon(\bar{x})$ and $\pi_x^k \in \mathbb{B}_\varepsilon(\bar{x})$ that

$$\|\pi_x^k - \hat{x}^k + e_k\| \leq \frac{1}{\alpha_k}\Big(\frac{L_2}{2}\operatorname{dist}(x^k; \mathcal{X})^2 + \|B_k - \nabla^2 f(x^k)\| \operatorname{dist}(x^k; \mathcal{X}) + \alpha_k \operatorname{dist}(x^k; \mathcal{X}) + (1+M)\nu\|r(x^k)\|^{1+\varrho}\Big).$$

Finally, we employ the relationships

$$\|d^k\| = \|\hat{x}^k - x^k\| \leq \|\pi_x^k - \hat{x}^k + e_k\| + \|\pi_x^k - x^k\| + \|e_k\| \ \text{ and } \ \|e_k\| \leq \nu\|r(x^k)\|^{1+\varrho} \tag{25}$$

to justify the claimed estimate of the lemma. □

Now we are ready to derive the main result of this section about fast local convergence of Algorithm 1 under the metric subregularity of the subdifferential.

Theorem 4.1. *Let $\{x^k\}$ be the sequence of iterates generated by Algorithm* 1 *with $\alpha_k = c\|r(x^k)\|^\rho$, $\rho \in (0,1]$, and $\varrho \geq \rho$, and let $\bar{x} \in \mathcal{X}$ be any limiting point of the sequence $\{x^k\}_{k \in K_0}$, where K_0 is defined in* (22). *In addition to the standing assumptions, suppose that the subgradient mapping $\nabla f(x) + \partial g(x)$ is metrically subregular at $(\bar{x}, 0)$, that $\nabla^2 f$ is locally Lipschitzian around $\bar{x}$, and that $\|B_k - \nabla^2 f(x^k)\| = O(\|r(x^k)\|)$. Then there exists a natural number k_0 such that we have*

$$t_k = 1 \ \textit{ for all } \ k \geq k_0,$$

and that the sequence $\{x^k\}$ converges to the point $\bar{x}$. Furthermore, this convergence is superlinear with the rate $1+\rho$, i.e., there exist a positive number C_0 and a natural number k_0 for which

$$\operatorname{dist}(x^{k+1};\mathcal{X}) \le C_0 \operatorname{dist}(x^k;\mathcal{X})^{1+\rho} \quad \textit{whenever} \quad k \ge k_0. \tag{26}$$

In particular, for $\rho = 1$ we have the quadratic convergence of $x^k \to \bar{x}$ with the exponent $1+\rho = 2$ in (26).

Proof. Observe first that the assumed metric subregularity of the mapping $\nabla f(x) + \partial g(x)$ at $(\bar{x}, 0)$ gives us a positive number ε_1 and a natural number κ_1 such that for all p near $0 \in \mathbb{R}^n$ we have the inclusion

$$\Sigma(p) \cap \mathbb{B}_{\varepsilon_1}(\bar{x}) \subset \mathcal{X} + \kappa_1 \|p\| \mathbb{B} \;\text{ with }\; \Sigma(p) := \big\{ x \in \mathbb{R}^n \;\big|\; p \in \nabla f(x) + \partial g(x) \big\}. \tag{27}$$

Employing Proposition 2.1 allows us to find $\kappa_2 > 1$ ensuring the estimate

$$\operatorname{dist}(x;\mathcal{X}) \le \kappa_2 \|r(x)\| \;\text{ whenever }\; x \in \mathbb{B}_{\varepsilon_1}(\bar{x}). \tag{28}$$

Since $\|B_k - \nabla^2 f(x^k)\| = O(\|r(x^k)\|)$, we deduce from Proposition 2.2 that there exists $C_1 > 0$ such that

$$\|B_k - \nabla^2 f(x^k)\| \le C_1 \operatorname{dist}(x^k;\mathcal{X}).$$

Since $\nabla^2 f$ is locally Lipschitzian around $\bar{x}$, there exist numbers $L_2, \varepsilon_2 > 0$ such that

$$\|\nabla f(x) + \nabla^2 f(x)(y-x) - \nabla f(y)\| \le \frac{L_2}{2}\|x-y\|^2 \;\text{ for any }\; x, y \in \mathbb{B}_{\varepsilon_2}(\bar{x}). \tag{29}$$

Applying Lemma 4.1, we can select $\varepsilon_1 \in (0,1)$ to be so small that when $x^k \in \mathbb{B}_{\varepsilon_1}(\bar{x})$, $\|r(x^k)\| \le 1$ and

$$\|d^k\| \le \frac{1}{\alpha_k}\Big(\frac{L_2}{2}\operatorname{dist}(x^k;\mathcal{X})^2 + \|B_k - \nabla^2 f(x^k)\|\operatorname{dist}(x^k;\mathcal{X}) + 2\alpha_k \operatorname{dist}(x^k;\mathcal{X}) + (1+M+\alpha_k)\nu\|r(x^k)\|^{1+\varrho}\Big)$$

$$\le \frac{1}{c}\left(\frac{\kappa_2^\rho L_2}{2}\operatorname{dist}(x^k;\mathcal{X})^{2-\rho} + \kappa_2^\rho C_1 \operatorname{dist}(x^k;\mathcal{X})^{2-\rho} + 2\operatorname{dist}(x^k;\mathcal{X}) + (2+M)\nu\kappa_2^\rho(2+L_1)^{1+\varrho}\operatorname{dist}(x^k;\mathcal{X})^{1+\varrho-\rho}\right),$$

where the second inequality follows from Proposition 2.2 and the estimates

$$\|B_k - \nabla^2 f(x^k)\| \le C_1 \operatorname{dist}(x^k;\mathcal{X}) \;\text{ and }\; \alpha_k = c\|r(x^k)\|^\rho \ge c\operatorname{dist}(x;\mathcal{X})^\rho/\kappa_2^\rho.$$

Remembering that $\rho \in (0,1]$ and $\varrho \ge \rho$, we find $c_1 > 0$ such that

$$\|d^k\| \le c_1 \operatorname{dist}(x^k;\mathcal{X}) \;\text{ for all }\; x^k \in \mathbb{B}_{\varepsilon_1}(\bar{x}). \tag{30}$$

Since $\hat{x}^k$ is an inexact solution of (3) satisfying (14), Lemma 3.1 gives us a vector e_k for which the conditions in (16) hold. By setting $\tilde{x}^k := \hat{x}^k - e_k$, we have the inclusion

$$e_k - H_k e_k \in \nabla f(x^k) + H_k(\tilde{x}^k - x^k) + \partial g(\tilde{x}^k),$$

which implies therefore the following one

$$\mathcal{R}_k(\tilde{x}^k\, x^k) := \nabla f(\tilde{x}^k) - \nabla f(x^k) - H_k(\tilde{x}^k - x^k) + e_k - H_k e_k \in \nabla f(\tilde{x}^k) + \partial g(\tilde{x}^k). \tag{31}$$

The latter reads, by the above definition of the perturbed solution map $\Sigma(p)$, that $\tilde{x}^k \in \Sigma(\mathcal{R}_k(\tilde{x}^k\, x^k))$. Without loss of generality, let ε_1 be so small that $\varepsilon_1 < \varepsilon_2/(3+c_1+L_1)$ and then get $\|\tilde{x}^k - \bar{x}\| \le \|x^k - \bar{x}\| + \|d^k\| + \|e_k\| \le (1+c_1+2+L_1)\|x^k - \bar{x}\| < \varepsilon_2$. This leads us to the relationships

$$\begin{aligned}
\|\mathcal{R}_k(\tilde{x}^k\, x^k)\| &= \|\nabla f(\tilde{x}^k) - \nabla f(x^k) - H_k(\tilde{x}^k - x^k) + e_k - H_k e_k\| \\
&= \|\nabla f(\tilde{x}^k) - \nabla f(x^k) - (B_k + \alpha_k I)(\tilde{x}^k - x^k) + e_k - H_k e_k\| \\
&\le \|\nabla f(\tilde{x}^k) - \nabla f(x^k) - \nabla^2 f(x^k)(\tilde{x}^k - x^k)\| + \|B_k - \nabla^2 f(x^k)\| \cdot \|\tilde{x}^k - x^k\| + \alpha_k\|\tilde{x}^k - x^k\| \\
&\quad + (1+M)\|e_k\| \\
&\le \frac{L_2}{2}\|\tilde{x}^k - x^k\|^2 + C_1\,\mathrm{dist}(x^k;\mathcal{X})\|\tilde{x}^k - x^k\| + c\|r(x^k)\|^\rho\|\tilde{x}^k - x^k\| + (1+M)\nu\|r(x^k)\|^{1+\varrho},
\end{aligned}$$

where the second inequality follows from (29), $\|B_k - \nabla^2 f(x^k)\| \le C_1\,\mathrm{dist}(x^k;\mathcal{X})$, $\alpha_k = c\|r(x^k)\|^\rho$, and $\|e_k\| \le \nu\|r(x^k)\|^{1+\varrho}$. Using now $\|e_k\| \le \nu\|r(x^k)\| \le \nu(2+L_1)\,\mathrm{dist}(x^k;\mathcal{X})$ and (30), we obtain

$$\|\tilde{x}^k - x^k\| \le \|\hat{x}^k - x^k\| + \|e^k\| = \|d^k\| + \|e_k\| \le (c_1 + \nu(2+L_1))\,\mathrm{dist}(x^k;\mathcal{X}) \ \text{ for all } \ x^k \in \mathbb{B}_{\varepsilon_1}(\bar{x}).$$

Then Proposition 2.2 and the assumption of $\varrho \ge \rho$ give us a constant $c_2 > 0$ such that

$$\|\mathcal{R}_k(\tilde{x}^k, x^k)\| \le c_2\,\mathrm{dist}(x^k;\mathcal{X})^{1+\rho} \ \text{ for all } \ x^k \in \mathbb{B}_{\varepsilon_1}(\bar{x}). \tag{32}$$

Taking further $0 < \varepsilon_3 < \varepsilon_1/(3+c_1+L_1)$ with $x^k \in \mathbb{B}_{\varepsilon_3}(\bar{x})$, we have

$$\|\tilde{x}^k - \bar{x}\| \le \|x^k - \bar{x}\| + \|d^k\| + \|e_k\| \le (1+c_1+2+L_1)\|x^k - \bar{x}\| < \varepsilon_1,$$

and thus $\tilde{x}^k \in \mathbb{B}_{\varepsilon_1}(\bar{x})$. It follows from the metric subregularity assumption (27) that $\tilde{x}^k \in \Sigma(\mathcal{R}_k(\tilde{x}^k, x^k))$, which yields for $x^k \in \mathbb{B}_{\varepsilon_3}(\bar{x})$ the estimates

$$\mathrm{dist}(\tilde{x}^k;\mathcal{X}) \le \kappa_1\|\mathcal{R}_k(\tilde{x}^k, x^k)\| \le \kappa_1 c_2\,\mathrm{dist}(x^k;\mathcal{X})^{1+\rho} \ \text{ and}$$

$$\begin{aligned}
\mathrm{dist}(\hat{x}^k;\mathcal{X}) \quad &\le \mathrm{dist}(\tilde{x}^k;\mathcal{X}) + \|e_k\| \le \kappa_1 c_2\,\mathrm{dist}(x^k;\mathcal{X})^{1+\rho} + \nu\|r(x^k)\|^{1+\varrho} \\
&\le (\kappa_1 c_2 + \nu(2+L_1)^{1+\varrho})\mathrm{dist}(x^k;\mathcal{X})^{1+\rho}.
\end{aligned} \tag{33}$$

Since $\rho > 0$, this allows us to find $0 < \varepsilon_0 < \varepsilon_3$ such that

$$\mathrm{dist}(\hat{x}^k;\mathcal{X}) \le \frac{\sigma}{(2+L_1)\kappa_2}\,\mathrm{dist}(x^k;\mathcal{X}) \ \text{ for } \ x^k \in \mathbb{B}_{\varepsilon_0}(\bar{x}). \tag{34}$$

Remembering that $C > F(x^0) \ge F_*$ and that F is continuous on $\mathrm{dom}\, F$, we select ε_0 to be so small that

$$\sup_{x \in \mathbb{B}_{\varepsilon_0}(\bar{x}) \cap \mathrm{dom}\, F} F(x) \le C. \tag{35}$$

Next we introduce the positive constants

$$\tilde{\sigma} := \frac{\sigma}{(2+L_1)\kappa_2} < 1 \ \text{ and } \ \tilde{\varepsilon} := \frac{1-\tilde{\sigma}}{1+c_1}\varepsilon_0$$

and show that if $x^{k_0} \in \mathbb{B}_{\tilde{\varepsilon}}(\bar{x})$ with some $k_0 \in K_0$, then for any $k \ge k_0$ we have

$$k \in K_0, \ t_k = 1, \ x^{k+1} = \hat{x}^k, \ \text{ and } \ x^{k+1} \in \mathbb{B}_{\varepsilon_0}(\bar{x}). \tag{36}$$

To verify (36), set first $k := k_0$ and deduce from $x^k \in \mathbb{B}_{\tilde{\varepsilon}}(\bar{x})$ that

$$\|\hat{x}^k - \bar{x}\| \le \|x^k - \bar{x}\| + \|d_k\| \le \|x^k - \bar{x}\| + c_1 \operatorname{dist}(x^k; \mathcal{X}) \le (1 + c_1)\|x^k - \bar{x}\| \le \varepsilon_0.$$

It follows from (28), (34), Proposition 2.2, and $k_0 \in K_0$ that

$$\begin{aligned}\|r(\hat{x}^k)\| &\le (2 + L_1)\operatorname{dist}(\hat{x}^k; \mathcal{X}) \le (2 + L_1)\frac{\sigma}{(2 + L_1)\kappa_2} \operatorname{dist}(x^k; \mathcal{X})\\ &\le \sigma\|r(x^k)\| = \sigma\vartheta_k.\end{aligned}$$

Observe also that (35) obviously yields $F(\hat{x}^k) \le C$. Then by Step 4 of Algorithm 1 we get $k + 1 \in K_0$, $t_k = 1$, $x^{k+1} = \hat{x}^k$, $\vartheta_{k+1} = r(x^{k+1})$, and $x^{k+1} \in \mathbb{B}_{\varepsilon_0}(\bar{x})$. To justify further (36) for any $k > k_0$, proceed by induction and suppose that for all $k - 1 \ge \ell \ge k_0$ we have

$$\ell + 1 \in K_0,\ t_\ell = 1,\ x^{\ell+1} = \hat{x}^\ell,\ \vartheta_{\ell+1} = r(x^{\ell+1}),\ x^{\ell+1} \in \mathbb{B}_{\varepsilon_0}(\bar{x}), \text{ and hence } \operatorname{dist}(x^{\ell+1}; \mathcal{X}) \le \tilde{\sigma} \operatorname{dist}(x^\ell; \mathcal{X}).$$

Then we readily arrive at the estimates

$$\|\hat{x}^k - x^{k_0}\| \le \sum_{\ell=k_0}^{k} \|d^\ell\| \le \sum_{\ell=k_0}^{k} c_1 \operatorname{dist}(x^\ell; \mathcal{X}) \le \sum_{\ell=k_0}^{k} c_1 \tilde{\sigma}^{\ell-k_0} \operatorname{dist}(x^{k_0}; \mathcal{X}) \le \frac{c_1}{1 - \tilde{\sigma}} \operatorname{dist}(x^{k_0}; \mathcal{X}) \le \frac{c_1}{1 - \tilde{\sigma}}\|x^{k_0} - \bar{x}\|, \tag{37}$$

where the second inequality follows from (30). This tells us that

$$\|\hat{x}^k - \bar{x}\| \le \|\hat{x}^k - x^{k_0}\| + \|x^{k_0} - \bar{x}\| \le \frac{1 + c_1}{1 - \tilde{\sigma}}\tilde{\varepsilon} \le \varepsilon_0.$$

Arguing as above, we get that $\|r(\hat{x}^k)\| \le \sigma\vartheta_k$ and $F(\hat{x}^k) \le C$, which ensures that (36) holds for $k + 1$ and thus verifies these conditions in the general case.

Now we prove the claimed convergence $x^k \to \bar{x}$ as $k \to \infty$ with the convergence rate (26), where $\bar{x}$ is the designated limiting point $\bar{x}$ of the sequence $\{x^k\}_{k \in K_0}$. Using conditions in (36) and arguments similarly to (37), we are able to show that, for any $\tilde{k} \in K_0$ with $\tilde{k} \ge k_0$,

$$\|x^k - \bar{x}\| \le \frac{c_1}{1 - \tilde{\sigma}}\|x^{\tilde{k}} - \bar{x}\| + \|x^{\tilde{k}} - \bar{x}\| \quad \text{whenever} \quad k \ge \tilde{k}. \tag{38}$$

This shows that the sequence $\{x^k\}$ is bounded. Picking any limiting point $\tilde{x}$ of $\{x^k\}$ and passing to the limit as $k \to \infty$ in (38) lead us to estimate

$$\|\tilde{x} - \bar{x}\| \le \frac{c_1}{1 - \tilde{\sigma}}\|x^{\tilde{k}} - \bar{x}\| + \|x^{\tilde{k}} - \bar{x}\|.$$

Recalling that $\bar{x}$ is a limiting point of $\{x^k\}_{k \in K_0}$, we pass to the limit as $\tilde{k} \to \infty$ in the estimate above and get $\|\tilde{x} - \bar{x}\| = 0$, which implies that $\{x^k\}$ converges to $\bar{x}$. Finally, employing (33) gives us numbers $C_0, k_0 > 0$ such that the claimed condition (26) holds. This completes the proof of the theorem. □

5 Fast Local Convergence under Metric q-Subregularity

In this section we study the local convergence of Algorithm 1 under the metric q-subregularity of the subgradient mapping in (1) in both cases where $q \in (0, 1]$ and $q > 1$. In the first case, referred to as the Hölder metric subregularity, we do not consider any $q \in (0, 1]$, but precisely specify the lower bound of q

and respectively modify some parameters of our algorithm. The imposed metric q-subregularity assumptions is weaker for $q < 1$ than the metric subregularity one required in Theorem 4.1, but nevertheless allows us to achieve a local superlinear (while not quadratic) convergence of the algorithm. In the other case where $q > 1$, we achieve a higher than quadratic rate of local convergence of the proposed algorithm.

Starting with the Hölder metric subregularity, we first provide the following direction estimate.

Lemma 5.1. *Let $\{x^k\}$ be the sequence generated by Algorithm 1 with $\alpha_k = c\|r(x^k)\|^\rho$ where $\rho \in (0,1]$ and $\varrho \geq 1$, and let $\bar{x} \in \mathcal{X}$ be any limiting point of $\{x^k\}$. In addition to the standing assumptions, suppose that the subgradient mapping $\nabla f(x) + \partial g(x)$ is metrically q-subregular at $(\bar{x}, 0)$ for some $q \in (0,1]$, that the Hessian $\nabla^2 f$ is locally Lipschitzian around $\bar{x}$, and that the estimate $\|B_k - \nabla^2 f(x^k)\| \leq C_1 \operatorname{dist}(x^k;\mathcal{X})$ holds with some constant $C_1 > 0$. Then there exist positive numbers ε and c_1 such that for $d^k := \hat{x}^k - x^k$ we have*

$$\alpha_k\|d^k\| \leq c_1 \operatorname{dist}(x^k;\mathcal{X})^{1+\rho} \quad \textit{and} \quad \|d^k\| \leq c_1 \max\left\{\operatorname{dist}(x^k;\mathcal{X})^{2-\frac{\varrho}{q}}, \operatorname{dist}(x^k;\mathcal{X})\right\} \quad \textit{as} \quad x^k \in \mathbb{B}_\varepsilon(\bar{x}). \tag{39}$$

Proof. Similarly to the proof of Lemma 4.1, observe that there exist $\varepsilon_0, L_2 > 0$ such that

$$\|d^k\| \leq \frac{1}{\alpha_k}\Big(\frac{L_2}{2}\operatorname{dist}(x^k;\mathcal{X})^2 + \|B_k - \nabla^2 f(x^k)\|\operatorname{dist}(x^k;\mathcal{X}) + 2\alpha_k \operatorname{dist}(x^k;\mathcal{X}) + (1+M+\alpha_k)\nu\|r(x^k)\|^{1+\varrho}\Big)$$

provided that $x^k \in \mathbb{B}_{\varepsilon_0}(\bar{x})$. Since $\|r(x^k)\| \leq (2+L_1)\operatorname{dist}(x^k;\mathcal{X})$ by Proposition 2.2, and since $\|B_k - \nabla^2 f(x^k)\| \leq C_1 \operatorname{dist}(x^k;\mathcal{X})$ by the imposed assumption, we have for such x^k that

$$\alpha_k\|d^k\| \leq \left(\frac{L_2}{2} + C_1\right)\operatorname{dist}(x^k;\mathcal{X})^2 + 2\alpha_k \operatorname{dist}(x^k;\mathcal{X}) + (1+M+\alpha_k)\nu(2+L_1)^{1+\varrho}\operatorname{dist}(x^k;\mathcal{X})^{1+\varrho}.$$

The assumed metric q-subregularity of $\nabla f(x) + \partial g$ gives us by Proposition 2.3 numbers $\varepsilon_1, \kappa_1 > 0$ with

$$\operatorname{dist}(x;\mathcal{X}) \leq \kappa_1\|r(x)\|^q \quad \text{for all} \quad x \in \mathbb{B}_{\varepsilon_1}(\bar{x}).$$

Remembering that $\alpha_k = c\|r(x^k)\|^\rho$ with $\rho \in (0,1]$ implies that

$$\alpha_k = c\|r(x^k)\|^\rho \geq c\kappa_1^{-\frac{\varrho}{q}}\operatorname{dist}(x;\mathcal{X})^{\frac{\varrho}{q}} \quad \text{as} \quad x^k \in \mathbb{B}_{\varepsilon_1}(\bar{x}). \tag{40}$$

Since $\rho \in 0,1]$ and $\varrho \geq 1$, we deduce from the latter the existence of a positive number c_1 ensuring the fulfillment of both estimates claimed in the lemma. □

Having Lemma 5.1 and some previous estimates in hand, next we derive the following superlinear convergence result for Algorithm 1 with a particular choice of parameters under the assumed Hölder metric subregularity with an appropriate factor q. Observe that neither the assumptions nor the conclusions of Theorem 5.1 reduce to those in Theorem 4.1 even the case where $q = 1$ in the theorem below. Its proof follows the lines in the proof of Theorem 4.1 with more involved estimates.

Theorem 5.1. *Let $\{x^k\}$ be the sequence generated by Algorithm 1 with $\alpha_k = c\|r(x^k)\|^\rho$, $\rho = \frac{\sqrt{5}-1}{2}$, and $\varrho \geq 1$, and let $\bar{x} \in \mathcal{X}$ be any limiting point of the sequence $\{x^k\}_{k \in K_0}$, where K_0 is defined in (22). In addition to the standing assumptions, suppose that the subgradient mapping $\nabla f(x) + \partial g(x)$ is metrically q-subregular at $(\bar{x}, 0)$ with $q \in (\frac{\sqrt{5}-1}{2}, 1]$, that the Hessian mapping $\nabla^2 f$ is locally Lipschitzian around $\bar{x}$, and that $\|B_k - \nabla^2 f(x^k)\| = O(\|r(x^k)\|)$. Then there exist a natural number k_0 such that $t_k = 1$ for all $k \geq k_0$, and that the sequence $\{x^k\}$ superlinearly converges to $\bar{x}$ on the sense that*

$$\operatorname{dist}(x^{k+1};\mathcal{X}) = o\big(\operatorname{dist}(x^k;\mathcal{X})\big) \quad \textit{whenever} \quad k \geq k_0. \tag{41}$$

Proof. It follows from the metric q-subregularity (6) of $\nabla f(x) + \partial g(x)$ at $(\bar{x}, 0)$ that there exist $\varepsilon_1, \kappa_1 > 0$ such that for any p near the origin of $\mathbb{R}^n$ we have

$$\Sigma(p) \cap \mathbb{B}_{\varepsilon_1}(\bar{x}) \subset \mathcal{X} + \kappa_1 \|p\|^q \mathbb{B} \tag{42}$$

for the solution map $\Sigma(p)$ of the perturbed generalized equation defined in the proof of Theorem 4.1. Proposition 2.3 gives us a constant $\kappa_2 > 1$ for which

$$\operatorname{dist}(x; \mathcal{X}) \leq \kappa_2 \|r(x)\|^q \ \text{ whenever } \ x \in \mathbb{B}_{\varepsilon_1}(\bar{x}). \tag{43}$$

Since $\|B_k - \nabla^2 f(x^k)\| = O(\|r(x^k)\|)$, we deduce from Proposition 2.2 the existence of $C_1 > 0$ with

$$\|B_k - \nabla^2 f(x^k)\| \leq C_1 \operatorname{dist}(x^k; \mathcal{X}). \tag{44}$$

Recalling that $\hat{x}^k$ is an inexact solution of (3) satisfying (14) and using Lemma 3.1 give us

$$e_k \in \nabla f(x^k) + H_k(\hat{x}^k - x^k) + \partial g(\hat{x}^k - e_k) \ \text{ with } \ \|e_k\| \leq \nu \|r(x^k)\|^{1+\varrho}.$$

Arguing as in the proof of Theorem 4.1, we get $\tilde{x}^k \in \Sigma(\mathcal{R}_k(\tilde{x}^k, x^k))$, where $\tilde{x}^k := \hat{x}^k - e_k$ and $\mathcal{R}_k(\tilde{x}^k, x^k)$ is defined in (31), and obtain (29) with some $L_2, \varepsilon_2 > 0$. Then choose by Lemma 5.1 a small $\varepsilon_1 > 0$ such that

$$\|d^k\| \leq c_1 \max\left\{\operatorname{dist}(x^k; \mathcal{X})^{2-\frac{\varrho}{q}}, \operatorname{dist}(x^k; \mathcal{X})\right\} \ \text{ for all } \ x^k \in \mathbb{B}_{\varepsilon_1}(\bar{x}) \tag{45}$$

with some $c_1 > 0$. Letting $\varepsilon_1 < \min\{1, \varepsilon_2\}$ and following the proof of Theorem 4.1, we arrive at the estimate

$$\|\mathcal{R}_k(\tilde{x}^k\, x^k)\| \leq \frac{L_2}{2}\|\tilde{x}^k - x^k\|^2 + C_1 \operatorname{dist}(x^k; \mathcal{X})\|\tilde{x}^k - x^k\| + \alpha_k \|\tilde{x}^k - x^k\| + (1+M)\nu\|r(x^k)\|^{1+\varrho} \tag{46}$$

if $x^k \in \mathbb{B}_{\varepsilon_1}(\bar{x})$. Since $\|e_k\| \leq \nu\|r(x^k)\| \leq \nu(2 + L_1)\operatorname{dist}(x^k; \mathcal{X})$ for this choice of x^k, it follows that

$$\|\tilde{x}^k - x^k\| \leq \|\hat{x}^k - x^k\| + \|e^k\| = \|d^k\| + \|e_k\| \leq \|d^k\| + \nu(2 + L_1)\operatorname{dist}(x^k; \mathcal{X}),$$

which being combined with (46), Proposition 2.2 and by taking into account that $\varrho \geq 1$ gives us $c_2 > 0$ with

$$\|\mathcal{R}_k(\tilde{x}^k\, x^k)\| \leq c_2\|d^k\|^2 + c_2 \operatorname{dist}(x^k; \mathcal{X})^2 + \alpha_k\left(\|d^k\| + \nu(2 + L_1)\operatorname{dist}(x^k; \mathcal{X})\right) \ \text{ for all } \ x^k \in \mathbb{B}_{\varepsilon_1}(\bar{x}).$$

Then the direction estimate (45) together with the one of $\alpha_k = c\|r(x^k)\|^\rho \leq c(2 + L_1)^\rho \operatorname{dist}(x^k; \mathcal{X})^\rho$, which comes from Proposition 2.2, ensures the existence of a constant $c_3 > 0$ providing the condition

$$\|\mathcal{R}_k(\tilde{x}^k\, x^k)\| \leq c_3 \max\left\{\operatorname{dist}(x^k; \mathcal{X})^{4-2\frac{\varrho}{q}}, \operatorname{dist}(x^k; \mathcal{X})^{1+\rho}\right\} \ \text{ for all } \ x^k \in \mathbb{B}_{\varepsilon_1}(\bar{x}).$$

Since $\|\tilde{x}^k - \bar{x}\| \leq \|x^k - \bar{x}\| + \|d^k\| + \|e_k\|$ with $\|d^k\| \to 0$ and $\|e_k\| \to 0$ when $x^k \to \bar{x}$ as $k \to \infty$, we find $0 < \varepsilon_3 \leq \varepsilon_1$ such that $\tilde{x}^k \in \mathbb{B}_{\varepsilon_1}(\bar{x})$ whenever $x^k \in \mathbb{B}_{\varepsilon_3}(\bar{x})$. Recalling that $\tilde{x}^k \in \Sigma(\mathcal{R}_k(\tilde{x}^k\, x^k))$ and then employing the metric q-subregularity condition (42) tell us that

$$\operatorname{dist}(\tilde{x}^k; \mathcal{X}) \leq \kappa_1\|\mathcal{R}_k(\tilde{x}^k, x^k)\|^q \leq \kappa_1 c_3 \max\left\{\operatorname{dist}(x^k; \mathcal{X})^{4q-2\rho}, \operatorname{dist}(x^k; \mathcal{X})^{(1+\rho)q}\right\} \ \text{ if } \ x^k \in \mathbb{B}_{\varepsilon_3}(\bar{x}).$$

Combining the latter with $\operatorname{dist}(\hat{x}^k; \mathcal{X}) \leq \operatorname{dist}(\tilde{x}^k; \mathcal{X}) + \|e_k\|$ and $\|e_k\| \leq \nu\|r(x^k)\|^{1+\varrho}$ as $\varrho \geq 1$, we have

$$\operatorname{dist}(\hat{x}^k; \mathcal{X}) \leq (\kappa_1 c_3 + \nu(2 + L_1)^{1+\varrho})\operatorname{dist}(x^k; \mathcal{X})^{\min\{4q-2\rho, (1+\rho)q, 2\}} \ \text{ when } \ x^k \in \mathbb{B}_{\varepsilon_3}(\bar{x}).$$

The choice of the Hölder metric subregularity parameter $q > \rho = \frac{-1+\sqrt{5}}{2}$ yields $4q - 2\rho > 1$ and $(1+\rho)q > 1$, and therefore it gives us the estimate

$$\operatorname{dist}(\hat{x}^k; \mathcal{X}) = o\big(\operatorname{dist}(x^k; \mathcal{X})\big) \;\text{ when }\; x^k \in \mathbb{B}_{\varepsilon_3}(\bar{x}). \tag{47}$$

This ensures that for any $\tilde{\sigma} \in (0, 1)$ there exists $0 < \varepsilon_0 < \varepsilon_3$ such that

$$\operatorname{dist}(\hat{x}^k; \mathcal{X}) \le \tilde{\sigma} \operatorname{dist}(x^k; \mathcal{X}) \;\text{ if }\; x^k \in \mathbb{B}_{\varepsilon_0}(\bar{x}).$$

It follows from (29) that whenever $x^k \in \mathbb{B}_{\varepsilon_0}(\bar{x})$ we have

$$\left| f(\hat{x}^k) - f(x^k) - \nabla f(x^k)^T d^k - \frac{1}{2}(d^k)^T \nabla^2 f(x^k)(d^k) \right| \le \frac{L_2}{2} \|d^k\|^3.$$

Using the definition of l_k in (2) and the above estimates brings us to the relationships

$$\begin{aligned}
F(x^k) - F(\hat{x}^k) - \frac{\theta \alpha_k}{2}\|d^k\|^2 &\ge l_k(x^k) - l_k(\hat{x}^k) - \frac{1}{2}(d^k)^T \nabla^2 f(x^k)(d^k) - \frac{L_2}{2}\|d^k\|^3 - \frac{\theta\alpha_k}{2}\|d^k\|^2 \\
&\ge \frac{1}{2}(d^k)^T H_k(d^k) - \frac{1}{2}(d^k)^T \nabla^2 f(x^k)(d^k) - \frac{L_2}{2}\|d^k\|^3 - \frac{\theta\alpha_k}{2}\|d^k\|^2 \\
&\ge \frac{\alpha_k(1-\theta)}{2}\|d^k\|^2 + \frac{1}{2}(d^k)^T \big(B_k - \nabla^2 f(x^k)\big)(d^k) - \frac{L_2}{2}\|d^k\|^3 \\
&\ge \frac{\alpha_k(1-\theta)}{2}\|d^k\|^2 - \frac{C}{2}\operatorname{dist}(x^k; \mathcal{X})\|d^k\|^2 - \frac{L_2}{2}\|d^k\|^3 \\
&\ge \frac{1}{2}\|d^k\|^2 \left((1-\theta)\alpha_k - C\operatorname{dist}(x^k; \mathcal{X}) - L_2\|d^k\|\right),
\end{aligned}$$

where the second inequality follows from (19) while the fourth inequality is a consequence of (44). By the estimates on d_k in (39) and α_k in (40) established in Lemma 5.1 and by the imposed condition $q > \rho$, we can choose $\varepsilon_0 > 0$ to be sufficiently small that

$$F(x^k) - F(\hat{x}^k) - \frac{\theta\alpha_k}{2}\|d^k\|^2 \ge 0 \;\text{ when }\; x^k \in \mathbb{B}_{\varepsilon_0}(\bar{x}).$$

Consider further the positive number

$$\tilde{\varepsilon} := \min\left\{ \frac{\varepsilon_0}{2}, \frac{\varepsilon_0}{2c_1}, \Big(\frac{1 - \tilde{\sigma}^{2-\rho/q}}{4mc_1}\varepsilon_0\Big)^{\frac{1}{2-\rho/q}}, \frac{1-\tilde{\sigma}}{4c_1}\varepsilon_0 \right\}$$

and show that for any $k \ge k_0$ we have $t_k = 1$, $x^{k+1} = \hat{x}^k$, and $x^{k+1} \in \mathbb{B}_{\varepsilon_0}(\bar{x})$ whenever $x^{k_0} \in \mathbb{B}_{\tilde{\varepsilon}}(\bar{x})$. Indeed, letting $k := k_0$ and remembering that $x^k \in \mathbb{B}_{\tilde{\varepsilon}}(\bar{x})$ tell us that

$$\begin{aligned}
\|\hat{x}^k - \bar{x}\| &\le \|x^k - \bar{x}\| + \|d_k\| \\
&\le \|x^k - \bar{x}\| + c_1 \max\left\{\operatorname{dist}(x^k; \mathcal{X})^{2-\frac{\rho}{q}}, \operatorname{dist}(x^k; \mathcal{X})\right\} \\
&\le \|x^k - \bar{x}\| + c_1 \max\left\{\|x^k - \bar{x}\|^{2-\frac{\rho}{q}}, \|x^k - \bar{x}\|\right\} \le \varepsilon_0,
\end{aligned}$$

where the second inequality follows from (45). In this case we have $t_k = 1$ from the Step 4 of Algorithm 1 provided that $\|r(\hat{x}^k)\| \le \sigma\vartheta_k$ and $f(\hat{x}^k) \le C$. Otherwise, it follows from $x^k \in \mathbb{B}_{\varepsilon_0}(\bar{x})$ that

$$F(\hat{x}^k) \le F(x^k) - \frac{\theta\alpha_k}{2}\|d^k\|^2,$$

and so $t_k = 1$ by Step 5 of the algorithm. Thus in both settings with $k = k_0$ we have $x^{k+1} = \hat{x}^k \in \mathbb{B}_{\varepsilon_0}(\bar{x})$.

In the remaining case where $k > k_0$, we proceed by induction similarly to the proof of Theorem 4.1 and assume that $t_\ell = 1$, $x^{\ell+1} = \hat{x}^\ell$, and $x^{\ell+1} \in \mathbb{B}_{\varepsilon_0}(\bar{x})$ for all $k-1 \geq \ell \geq k_0$. This implies that $\operatorname{dist}(x^{\ell+1}; \mathcal{X}) \leq \tilde{\sigma} \operatorname{dist}(x^\ell; \mathcal{X})$, and therefore we get the estimates

$$\begin{aligned}
\|\hat{x}^k - x^{k_0}\| &\leq \sum_{\ell=k_0}^{k} \|d^\ell\| \leq \sum_{\ell=k_0}^{k} c_1 \max\left\{\operatorname{dist}(x^\ell; \mathcal{X})^{2-\frac{\varrho}{q}}, \operatorname{dist}(x^\ell; \mathcal{X})\right\} \\
&\leq \sum_{\ell=k_0}^{k} c_1 \big(\operatorname{dist}(x^\ell; \mathcal{X})^{2-\frac{\varrho}{q}} + \operatorname{dist}(x^\ell; \mathcal{X})\big) \\
&\leq \sum_{\ell=k_0}^{k} c_1 \big(\tilde{\sigma}^{(2-\rho/q)(\ell-k_0)} \operatorname{dist}(x^{k_0}; \mathcal{X})^{2-\rho/q} + \tilde{\sigma}^{\ell-k_0} \operatorname{dist}(x^{k_0}; \mathcal{X})\big) \\
&\leq \frac{c_1}{1-\tilde{\sigma}^{2-\rho/q}} \operatorname{dist}(x^{k_0}; \mathcal{X})^{2-\rho/q} + \frac{c_1}{1-\tilde{\sigma}} \operatorname{dist}(x^{k_0}; \mathcal{X}) \\
&\leq \frac{c_1}{1-\tilde{\sigma}^{2-\rho/q}} \|x^{k_0} - \bar{x}\|^{2-\rho/q} + \frac{c_1}{1-\tilde{\sigma}} \|x^{k_0} - \bar{x}\| \leq \frac{\varepsilon_0}{2},
\end{aligned} \tag{48}$$

where the second inequality follows from (45) while the fifth one is due to $2 - \rho/q > 1$. Then

$$\|\hat{x}^k - \bar{x}\| \leq \|\hat{x}^k - x^{k_0}\| + \|x^{k_0} - \bar{x}\| \leq \varepsilon_0,$$

which allows us to justify that $t_k = 1$, $x^{k+1} = \hat{x}^k$, and $x^{k+1} \in \mathbb{B}_{\varepsilon_0}(\bar{x})$ by using the arguments that are similar to the case of $k = k_0$ furnished above.

To verify now the superlinear convergence statement (41) of the theorem, take the chosen limiting point $\bar{x}$ of the sequence $\{x^k\}$ and find $k_0 > 0$ such that $x^{k_0} \in \mathbb{B}_{\bar{\varepsilon}}(\bar{x})$. Then, as shown above, for any $k \geq k_0$ we have $t_k = 1$, $x^{k+1} = \hat{x}^k$, and $x^{k+1} \in \mathbb{B}_{\varepsilon_0}(\bar{x})$. Proceeding similarly to the proof of (48) leads us to the inequality

$$\|x^k - \bar{x}\| \leq \frac{c_1}{1-\tilde{\sigma}^{2-\rho/q}} \|x^{\tilde{k}} - \bar{x}\|^{2-\rho/q} + \frac{c_1}{1-\tilde{\sigma}} \|x^{\tilde{k}} - \bar{x}\| + \|x^{\tilde{k}} - \bar{x}\| \ \text{ for any } \ k > \tilde{k} \geq k_0. \tag{49}$$

Denote by $\tilde{x}$ an arbitrary limiting point of $\{x^k\}$ and fix any $\tilde{k} \geq k_0$ in (49). The passage to the limit in (49) as $k \to \infty$ gives us the estimate

$$\|\tilde{x} - \bar{x}\| \leq \frac{c_1}{1-\tilde{\sigma}^{2-\rho/q}} \|x^{\tilde{k}} - \bar{x}\|^{2-\rho/q} + \frac{c_1}{1-\tilde{\sigma}} \|x^{\tilde{k}} - \bar{x}\| + \|x^{\tilde{k}} - \bar{x}\|.$$

Then passing to the limit therein as $\tilde{k} \to \infty$, we arrive at $\|\tilde{x} - \bar{x}\| = 0$, which readily justifies that $\{x^k\}$ converges to $\bar{x}$. Using finally (47) brings us to (41) and thus completes the proof of the theorem. □

The final result of this section concerns the other kind of metric q-subregularity of the subgradient mapping in (1) in the case where $q > 1$. As discussed in Section 2, this type of higher-order metric subregularity is rather new in the literature, and it has never been used in applications to numerical optimization. The following theorem shows that the higher-order subregularity assumption imposed on the subgradient mapping ∂F at the point in question allows us to derive an extension of Theorem 4.1 with establishing the *convergence rate*, which may be *higher than quadratic*.

Theorem 5.2. *Let $\{x^k\}$ be the sequence generated by Algorithm 1 with $\alpha_k = c\|r(x^k)\|^\rho$ as $\rho \in (0, 1]$, and let $\bar{x} \in \mathcal{X}$ be any limiting point of $\{x^k\}_{k \in K_0}$, where the set K_0 is taken from (22). In addition to the standing assumptions, suppose that the mapping $\nabla f(x) + \partial g(x)$ is metrically q-subregular at $(\bar{x}, 0)$ with $q > 1$, that the Hessian $\nabla^2 f$ is locally Lipschitzian around $\bar{x}$, that $\|B_k - \nabla^2 f(x^k)\| = O(\|r(x^k)\|)$, and that $\varrho \geq q(1+\rho) - 1$*

in (14)*. Then there exists an index k_0 such that $t_k = 1$ for all $k \geq k_0$ and that the sequence $\{x^k\}$ converges to the point $\bar{x}$ with the convergence rate $q(1+\rho)$. The latter means that for some $k_0, C_0 > 0$ we have*

$$\operatorname{dist}(x^{k+1}; \mathcal{X}) \leq C_0 \operatorname{dist}(x^k; \mathcal{X})^{q(1+\rho)} \text{ whenever } k \geq k_0. \tag{50}$$

Proof. It follows from the imposed metric q-subregularity condition with a fixed degree $q > 1$ that

$$\Sigma(p) \cap \mathbb{B}_{\varepsilon_1}(\bar{x}) \subset \mathcal{X} + \kappa_1 \|p\|^q \mathbb{B} \text{ for some } \varepsilon_1, \kappa_1 > 0 \tag{51}$$

whenever $p \in \mathbb{R}^n$ is sufficiently close to the origin. Following the proof of Theorem 4.1, we arrive at the estimate of $\|\mathcal{R}_k(\tilde{x}^k, x^k)\|$ in (32) with some constant $c_2 > 0$, where $\tilde{x}^k := \hat{x}^k - e_k$ while $\mathcal{R}_k(\tilde{x}^k, x^k)$, $\hat{x}^k$, and e_k are defined and analyzed similarly to the case of Theorem 4.1. Then there exists $\varepsilon_3 > 0$ such that $\tilde{x}^k \in \mathbb{B}_{\varepsilon_1}(\bar{x})$ when $x^k \in \mathbb{B}_{\varepsilon_3}(\bar{x})$. Since $\tilde{x}^k \in \Sigma(\mathcal{R}_k(\tilde{x}^k x^k))$, we combine this with (51) and get the estimates

$$\operatorname{dist}(\tilde{x}^k; \mathcal{X}) \leq \kappa_1 \|\mathcal{R}_k(\tilde{x}^k, x^k)\|^q \leq \kappa_1 c_2^q \operatorname{dist}(x^k; \mathcal{X})^{q(1+\rho)} \text{ and}$$

$$\begin{aligned} \operatorname{dist}(\hat{x}^k; \mathcal{X}) &\leq \operatorname{dist}(\tilde{x}^k; \mathcal{X}) + \|e_k\| \leq \kappa_1 c_2^q \operatorname{dist}(x^k; \mathcal{X})^{q(1+\rho)} + \nu \|r(x^k)\|^{1+\varrho} \\ &\leq (\kappa_1 c_2 + \nu(2+L_1)^{1+\varrho}) \operatorname{dist}(x^k; \mathcal{X})^{q(1+\rho)} \text{ whenever } x^k \in \mathbb{B}_{\varepsilon_3}(\bar{x}). \end{aligned} \tag{52}$$

Employing the induction arguments as in the proof of Theorem 4.1 yields the existence of a natural number k_0 such that we have $t_k = 1$, $x^{k+1} = \hat{x}^k$, $x^{k+1} \in \mathbb{B}_{\varepsilon_3}(\bar{x})$ when $k \geq k_0$, and that the sequence $\{x^k\}$ converges to $\bar{x}$ as $k \to \infty$. Hence the second estimate in (52) gives a positive number C_0 and a natural number k_0, which ensure the fulfillment the claimed convergence rate (50) and thus complete the proof of the theorem. □

6 Superlinear Local Convergence with Non-Lipschitzian Hessians

As seen in Sections 4 and 5, the imposed local Lipschitz continuity of the Hessian maping $\nabla^2 f$ plays a crucial role in the justifications of the local convergence results obtained therein. In this section we show that the latter assumption can be dropped with preserving a local superlinear convergence of Algorithm 1 for a rather broad class of loss functions f that naturally appear in many practical models arising in machine learning and statistics, which includes, e.g., linear regression, logistic regression, and Poisson regression.

The class of loss functions f of our consideration in this section satisfies the following structural properties.

Assumption 6.1. *The loss function $f\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ of* (1) *is represented in the form*

$$f(x) := h(Ax) + \langle b, x \rangle, \tag{53}$$

where A is an $m \times n$ matrix, $b \in \mathbb{R}^n$, and $h\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ is a proper, convex, and l.s.c. function such that:

- **(i)** *h is strongly convex on any compact and convex subset of the domain* $\operatorname{dom} h$*.*
- **(ii)** *h is continuously differentiable on the set* $\operatorname{dom} h$*, which is assumed to be open, and the gradient mapping ∇h is Lipschitz continuous on any compact subset $\Omega \subset$* $\operatorname{dom} h$*.*

Due to the strong convexity of h, the linear mapping $x \to Ax$ in (53) is invariant over the solution set $\mathcal{X}$ to (1). This is the contents of the following result taken from [23, Lemma 2.1].

Lemma 6.1. *Under the fulfillment of Assumption* 6.1 *there exists $\bar{y} \in \mathbb{R}^m$ such that $Ax = \bar{y}$ for all $x \in \mathcal{X}$.*

The next lemma is a counterpart of Lemma 4.1 without imposition the local Lipschitz continuity of the Hessian $\nabla^2 f$. By furnishing a similar while somewhat different scheme in comparison with Lemma 4.1, we establish new direction estimates of Algorithm 1 used in what follows. Note that we do not exploit in the lemma the structural conditions on f listed in Assumption 6.1.

Lemma 6.2. *Let $\{x^k\}$ be the sequence generated by Algorithm* 1 *with $\alpha_k = c\|r(x^k)\|^\rho$ and $\rho \in (0,1)$, and let $\bar{x} \in \mathcal{X}$ be any limiting point of $\{x^k\}$. In addition to Assumption* 1.1 *and* (10)*, suppose that the Hessian mapping $\nabla^2 f$ is continuous at $\bar{x} \in \mathcal{X}$, that $\|B_k - \nabla^2 f(x^k)\| \to 0$ as $k \to \infty$, and that the subgradient mapping $\nabla f(x) + \partial g(x)$ is metrically subregular at $(\bar{x}, 0)$. Then given an arbitrary quantity $\delta > 0$, there exist a positive number ε and a natural number k_0 such that for $d^k := \hat{x}^k - x^k$ we have the estimates*

$$\alpha_k \|d^k\| \le \delta \operatorname{dist}(x^k; \mathcal{X}) \quad \textit{and} \quad \|d^k\| \le \delta \operatorname{dist}(x^k; \mathcal{X})^{1-\rho} \quad \textit{when} \quad x^k \in \mathbb{B}_\varepsilon(\bar{x}) \quad \textit{and} \quad k > k_0. \tag{54}$$

Proof. Since $\hat{x}^k$ is an inexact solution to subproblem (3) satisfying (14), we get by Lemma 3.1 that there exists e_k for which both conditions in (16) hold. Taking the projection π_x^k of x^k onto the solution set $\mathcal{X}$ and arguing as in the proof of Lemma 3.1 bring us to the inequality in (24), which together with (54) and the direction estimate in (25) ensures that

$$\begin{aligned} \alpha_k \|d^k\| \le &\Big(\|\nabla f(x^k) + \nabla^2 f(x^k)(\pi_x^k - x^k) - \nabla f(\pi_x^k)\| + \|B_k - \nabla^2 f(x^k)\| \operatorname{dist}(x^k; \mathcal{X}) \\ &+ 2\alpha_k \operatorname{dist}(x^k; \mathcal{X}) + (1 + M + \alpha_k)\nu \|r(x^k)\|^{1+\varrho} \Big). \end{aligned} \tag{55}$$

It follows from the mean value theorem and the choice of π_x^k as the projection of x^k onto $\mathcal{X}$ that

$$\begin{aligned} \|\nabla f(x^k) + \nabla^2 f(x^k)(\pi_x^k - x^k) - \nabla f(\pi_x^k)\| &= \|(\nabla^2 f(x^k) - \nabla^2 f(\xi^k))(\pi_x^k - x^k)\| \\ &\le \|\nabla^2 f(x^k) - \nabla^2 f(\xi^k)\| \operatorname{dist}(x^k; \mathcal{X}), \end{aligned}$$

where $\xi^k := \lambda^k x^k + (1 - \lambda^k)\pi_x^k$ for some $\lambda^k \in (0,1)$, and hence $\xi^k \to \bar{x}$ when $x^k \to \bar{x}$ as $k \to \infty$. Then passing to the limit as $k \to \infty$ and using the assumed continuity of $\nabla^2 f$ at $\bar{x}$ show that $\|\nabla^2 f(x^k) - \nabla^2 f(\xi^k)\| \to 0$. Since $\alpha_k = c\|r(x^k)\|^\rho \to 0$ and $\|r(x^k)\| \le (2 + L_1)\operatorname{dist}(x^k; \mathcal{X})$ by Proposition 2.2, and since $\|B_k - \nabla^2 f(x^k)\| \to 0$ as $k \to \infty$, for any $\delta > 0$ we find a positive number ε and a natural number k_0 such that

$$\alpha_k \|d^k\| \le \delta \operatorname{dist}(x^k; \mathcal{X}) \quad \text{when} \quad x^k \in \mathbb{B}_\varepsilon(\bar{x}) \quad \text{and} \quad k > k_0,$$

which justifies the first estimate in (54). To verify finally the second one in (54), employ Proposition 2.1 and the above expression of α_k to find positive numbers ε_1 and c_1 ensuring the inequality

$$\alpha_k \ge c_1 \operatorname{dist}(x^k; \mathcal{X})^\rho \quad \text{for all} \quad x \in \mathbb{B}_{\varepsilon_1}(\bar{x}).$$

Combining the latter with the first estimate in (54) tells us that for the fixed number $\delta > 0$ there exist $\varepsilon > 0$ and $k > k_0$ such that the second estimate in (54) is also satisfied, and thus the proof is complete. □

Now we are ready to derive the promised result showing that the sequence of iterates, which are generated by Algorithm 1 for the structural problem (1) considered in this section, converges superlinearly to a given optimal solution $\bar{x} \in \mathcal{X}$ without the local Lipschitz continuity assumption on the Hessian mapping $\nabla^2 f$.

Theorem 6.1. *Let $\{x^k\}$ be the sequence of iterates generated by Algorithm* 1 *with $\alpha_k = \|r(x^k)\|^\rho$ and $\rho \in (0,1)$, and let $\bar{x} \in \mathcal{X}$ be any limiting point of the sequence $\{x^k\}_{k \in K_0}$ with the set K_0 defined in* (22)*.*

Suppose in addition to the assumptions of Lemma 6.2 *that the loss function f is given in the structural form* (53) *under the fulfillment of Assumption* 6.1*, and that at each iteration step k the matrix B_k is represented in the form $B_k = A^T D_k A$, where A is taken from* (53) *while $D_k \in \mathbb{R}^{m\times m}$ is some positive semidefinite matrix. Then there exists a natural number k_0 such that $t_k = 1$ for all $k \geq k_0$, and that the sequence $\{x^k\}$ converges to $\bar{x}$ with the superlinear convergence rate, i.e., there is k_1 for which we have*

$$\mathrm{dist}(x^{k+1};\mathcal{X}) = o\big(\mathrm{dist}(x^k;\mathcal{X})\big) \quad \textit{whenever} \quad k \geq k_1. \tag{56}$$

Proof. Proceeding similarly to the proof of Theorem 4.1, at each iteration step k we have the vector $\mathcal{R}_k(\tilde{x}^k, x^k)$ defined in (31) with $\tilde{x}^k := \hat{x}^k - e_k$, where $\hat{x}^k$ is an inexact solution of (3) satisfying (14), and where e_k is taken from (16). These relationships and the mean value theorem applied to the gradient mapping ∇f on $[x^k, \tilde{x}^k]$ give us a vector $\tilde{\xi}^k := \tilde{\lambda}^k x^k + (1-\tilde{\lambda}^k)\tilde{x}^k$ with some $\tilde{\lambda}^k \in (0,1)$ such that

$$\begin{aligned}
\|\mathcal{R}_k(\tilde{x}^k\, x^k)\| &= \|\nabla f(\tilde{x}^k) - \nabla f(x^k) - H_k(\tilde{x}^k - x^k) + e_k - H_k e_k\| \\
&= \|\nabla f(\tilde{x}^k) - \nabla f(x^k) - (B_k + \alpha_k I)(\tilde{x}^k - x^k) + e_k - H_k e_k\| \\
&\leq \|\nabla f(\tilde{x}^k) - \nabla f(x^k) - \nabla^2 f(x^k)(\tilde{x}^k - x^k)\| + \|B_k - \nabla^2 f(x^k)\| \cdot \|\tilde{x}^k - x^k\| \\
&\quad + \alpha_k \|\tilde{x}^k - x^k\| + (1+M)\|e_k\| \\
&\leq \|(\nabla^2 f(\tilde{\xi}^k) - \nabla^2 f(x^k))(\tilde{x}^k - x^k)\| + \|(B_k - \nabla^2 f(x^k))(\tilde{x}^k - x^k)\| \\
&\quad + \alpha_k \|d^k\| + (1+M)\nu\|r(x^k)\|^{1+\varrho}.
\end{aligned}$$

Let $\tilde{\pi}_x^k$ and π_x^k be the projections of $\tilde{x}^k$ and x^k onto $\mathcal{X}$, respectively. Then it follows from Lemma 6.1 that $A\tilde{\pi}_x^k = A\pi_x^k$. By Assumption 6.1 we have $\nabla^2 f(x) = A^T \nabla^2 h(x) A$, and thus

$$\big(\nabla^2 f(\tilde{\xi}^k) - \nabla^2 f(x^k)\big)(\tilde{x}^k - x^k) = \big(\nabla^2 f(\tilde{\xi}^k) - \nabla^2 f(x^k)\big)(\tilde{x}^k - \tilde{\pi}_x^k + \pi_x^k - x^k).$$

Using the assumed representation $B_k = A^T D_k A$ of the matrix B_k, we similarly get that

$$\big(B_k - \nabla^2 f(x^k)\big)(\tilde{x}^k - x^k) = \big(B_k - \nabla^2 f(x^k)\big)(\tilde{x}^k - \tilde{\pi}_x^k + \pi_x^k - x^k).$$

Plugging the obtained expressions into the above estimate of $\|\mathcal{R}_k\|$ gives us

$$\begin{aligned}
\|\mathcal{R}_k(\tilde{x}^k\, x^k)\| &\leq \|(\nabla^2 f(\tilde{\xi}^k) - \nabla^2 f(x^k))(\tilde{x}^k - \tilde{\pi}_x^k + \pi_x^k - x^k)\| + \|(B_k - \nabla^2 f(x^k))(\tilde{x}^k - \tilde{\pi}_x^k + pi_x^k - x^k)\| \\
&\quad + \alpha_k\|d^k\| + (1+M)\nu\|r(x^k)\|^{1+\varrho} \\
&\leq \|\nabla^2 f(\tilde{\xi}^k) - \nabla^2 f(x^k)\|\big(\mathrm{dist}(\tilde{x}^k;\mathcal{X}) + \mathrm{dist}(x^k;\mathcal{X}) \\
&\quad + \|B_k - \nabla^2 f(x^k)\|\big(\mathrm{dist}(\tilde{x}^k;\mathcal{X}) + \mathrm{dist}(x^k;\mathcal{X})\big) + \alpha_k\|d^k\| + (1+M)\nu(2+L_1)^{\varrho}\mathrm{dist}(x^k;\mathcal{X})^{1+\varrho}.
\end{aligned}$$

It follows from the second estimate of Lemma 6.2 that $\|d^k\| \to 0$ as $k \to \infty$ and $x^k \to \bar{x}$. Since $x^k \to \bar{x}$ implies that $\tilde{x}^k \to \bar{x}$ as $k \to \infty$. Then the assumed continuity of $\nabla^2 f$ at $\bar{x}$ and the above construction of $\tilde{\xi}^k$ tell us that $\|\nabla^2 f(\tilde{\xi}^k) - \nabla^2 f(x^k)\| \to 0$ as $k \to \infty$ and $x^k \to \bar{x}$. Now the first estimate of Lemma 6.2 ensures that $\alpha_k\|d^k\| = o(\mathrm{dist}(x^k;\mathcal{X}))$ as $k \to \infty$ and $x^k \to \bar{x}$. Combining this with $\|B_k - \nabla^2 f(x^k)\| \to 0$ as $k \to \infty$ allows us to conclude that for any $\delta > 0$ there exist a positive number ε and a natural number k_0 such that

$$\|\mathcal{R}_k(\tilde{x}^k, x^k)\| \leq \delta\left(\mathrm{dist}(\tilde{x}^k;\mathcal{X}) + \mathrm{dist}(x^k;\mathcal{X})\right) \quad \text{whenever} \quad x^k \in \mathbb{B}_\varepsilon(\bar{x}) \quad \text{and} \quad k > k_0. \tag{57}$$

It follows from the metric subregularity assumption of the theorem that we have inclusion (27) with the perturbed solution map $\Sigma(p)$ therein. Since $\tilde{x}^k \in \Sigma(\mathcal{R}_k(\tilde{x}^k, x^k))$ as shown above, there exists $\kappa_1 > 0$ with

$$\mathrm{dist}(\tilde{x}^k;\mathcal{X}) \leq \kappa_1\|\mathcal{R}_k(\tilde{x}^k, x^k)\| \leq \kappa_1\delta\left(\mathrm{dist}(\tilde{x}^k;\mathcal{X}) + \mathrm{dist}(x^k;\mathcal{X})\right) \quad \text{for all} \quad x^k \in \mathbb{B}_\varepsilon(\bar{x}) \quad \text{and} \quad k > k_0,$$

which implies that $\text{dist}(\tilde{x}^k;\mathcal{X}) = o(\text{dist}(x^k;\mathcal{X}))$ as $k \to \infty$. Recalling the estimates

$$\text{dist}(\hat{x}^k;\mathcal{X}) \le \text{dist}(\tilde{x}^k;\mathcal{X}) + \|e_k\| \;\text{ and }\; \|e_k\| \le \nu\|r(x^k)\|^{1+\varrho} \le \nu(2+L_1)^{1+\varrho}\text{dist}(x^k;\mathcal{X})^{1+\varrho} = o\big(\text{dist}(x^k;\mathcal{X})\big),$$

we readily get, for all the numbers $\delta, \varepsilon, k_0, k$ taken from (57), the conditions

$$\text{dist}(\hat{x}^k;\mathcal{X}) = o\big(\text{dist}(x^k;\mathcal{X})\big) \;\text{ and }\; \text{dist}(\hat{x}^k;\mathcal{X}) \le \delta\,\text{dist}(x^k;\mathcal{X}), \tag{58}$$

which ensure therefore the existence of positive numbers ε_0 and κ_2 such that

$$\text{dist}(\hat{x}^k;\mathcal{X}) \le \frac{\sigma}{(2+L_1)\kappa_2}\text{dist}(x^k;\mathcal{X}) \;\text{ whenever }\; x^k \in \mathbb{B}_{\varepsilon_0}(\bar{x}) \;\text{ and }\; k > k_0. \tag{59}$$

Employing Lemma 6.2, suppose without loss of generality that there exists $c_1 > 0$ with

$$\|d^k\| \le c_1\,\text{dist}(x^k;\mathcal{X})^{1-\rho} \;\text{ for all }\; x^k \in \mathbb{B}_{\varepsilon_0}(\bar{x}) \;\text{ and }\; k > k_0. \tag{60}$$

Since $C > F(x^0) \ge F_*$ in our algorithm, and since F is continuous on $\text{dom}\,F$, let pick $\varepsilon_0 > 0$ to be so small that condition (35) holds. Defining the positive numbers

$$\tilde{\sigma} := \frac{\sigma}{(2+L_1)\kappa_2} < 1 \;\text{ and }\; \tilde{\varepsilon} := \min\left\{\frac{\varepsilon_0}{2}, \Big(\frac{1-\tilde{\sigma}^{1-\rho}}{2c_1}\varepsilon_0\Big)^{\frac{1}{1-\rho}}\right\} \tag{61}$$

and invoking the set K_0 from (22), we intend to show that if $x^{k_1} \in \mathbb{B}_{\tilde{\varepsilon}}(\bar{x})$ with $k_1 > k_0$ and $k_1 \in K_0$, then

$$k+1 \in K_0,\; t_k = 1,\; x^{k+1} = \hat{x}^k, \;\text{ and }\; x^{k+1} \in \mathbb{B}_{\varepsilon_0}(\bar{x}) \;\text{ whenever }\; k \ge k_1. \tag{62}$$

To prove it by induction, observe first that for $k := k_1$ all the conditions in (62) can be verified similarly to the proof of (49) in Theorem 4.1 with the replacement of k_0 by k_1 therein. Considering now the general case where $k > k_1$ in (62), suppose that the latter holds for any $k-1 \ge \ell \ge k_1$, which clearly yields $\text{dist}(x^{\ell+1};\mathcal{X}) \le \tilde{\sigma}\,\text{dist}(x^\ell;\mathcal{X})$. Then the above estimates and the choice of the parameters in (61) ensure that

$$\begin{aligned}\|\hat{x}^k - x^{k_1}\| &\le \sum_{\ell=k_1}^{k}\|d^\ell\| \le \sum_{\ell=k_1}^{k} c_1\,\text{dist}(x^\ell;\mathcal{X})^{1-\rho}\\ &\le \sum_{\ell=k_1}^{k} c_1\tilde{\sigma}^{(1-\rho)(\ell-k_1)}\text{dist}(x^{k_1};\mathcal{X})^{1-\rho} \le \frac{c_1}{1-\tilde{\sigma}^{1-\rho}}\text{dist}(x^{k_1};\mathcal{X})^{1-\rho}\\ &\le \frac{c_1}{1-\tilde{\sigma}^{1-\rho}}\|x^{k_1}-\bar{x}\|^{1-\rho},\end{aligned} \tag{63}$$

where the second inequality follows from (60). Thus by (61) and (63) we have

$$\|\hat{x}^k - \bar{x}\| \le \|\hat{x}^k - x^{k_1}\| + \|x^{k_1} - \bar{x}\| \le \frac{c_1}{1-\tilde{\sigma}^{1-\rho}}\|x^{k_1}-\bar{x}\|^{1-\rho} + \|x^{k_1}-\bar{x}\| \le \varepsilon_0,$$

which readily implies, similarly to the case where $k = k_1$, the fulfillment of (62) for any $k \ge k_1$. Furthermore, remembering that $\bar{x}$ is a limiting point of $\{x^k\}_{k\in K_0}$ and using (62) together with (63) allow us to check that for any $\tilde{k} \in K_0$ with $\tilde{k} \ge k_1$ we have

$$\|x^k - \bar{x}\| \le \frac{c_1}{1-\tilde{\sigma}^{1-\rho}}\|x^{\tilde{k}}-\bar{x}\|^{1-\rho} + \|x^{\tilde{k}}-\bar{x}\| \;\text{ whenever }\; k > \tilde{k}.$$

Further, let $\tilde{x}$ be any limiting point of the original iterative sequence $\{x^k\}$. Then the passage to the limit in the above inequality as $k \to \infty$ gives us

$$\|\tilde{x} - \bar{x}\| \le \frac{c_1}{1 - \tilde{\sigma}^{1-\rho}} \|x^{\tilde{k}} - \bar{x}\|^{1-\rho} + \|x^{\tilde{k}} - \bar{x}\| \text{ for all } \tilde{k} \ge k_1.$$

Passing finally the limit as $\tilde{k} \to \infty$ in the latter inequality and recalling that $\bar{x}$ is a limiting point of $\{x^k\}_{k \in K_0}$ tell us that $\|\tilde{x} - \bar{x}\| = 0$, which verifies therefore that $\{x^k\}$ converges to $\bar{x}$ as $k \to \infty$. The claimed estimate (56) of the convergence rate follows now from (58), and this completes the proof of the theorem. □

To conclude this section, observe that the standard choice of $B_k = \nabla^2 f(x^k)$ in Algorithm 1 clearly implies that the assumed representation $B_k = A^T D_k A$ and the condition $\|B_k - \nabla^2 f(x^k)\| \to 0$ as $k \to \infty$ *hold automatically* due to $\nabla^2 f(x^k) = A^T \nabla^2 h(Ax^k) A$ and the positive semidefiniteness of the Hessian $\nabla^2 h(Ax^k)$ under Assumption 6.1 on the loss function f imposed here. Furthermore, observe from the proof of Theorem 6.1 in comparison with those of Theorems 5.1 and 5.2 that the corresponding counterparts of the latter results can be derived for structural problems of the type considered in this section under the *metric q-subregularity* property of the subgradient mapping $\nabla f(x) + \partial g(x)$ at $(\bar{x}, 0)$ combined with the other assumptions of Theorem 6.1 while without the local Lipschitz continuity of the Hessian mapping $\nabla^2 f$.

7 Numerical Experiments for Regularized Logistic Regression

In the last section of the paper we conduct numerical experiments on solving the l_1 regularized logistic regression problem to support our theoretical results and compare them with the numerical algorithm from [39] applicable to this problem. All the numerical experiments are implemented on a laptop with Intel(R) Core(TM) i7-9750H CPU@ 2.60GHz and 32.00 GB memory. All the codes are written in MATLAB 2019b.

Supposing we are given some training data pairs $(a_i, b_i) \in \mathbb{R}^n \times \{-1, 1\}$ as $i = 1, \ldots, N$, the optimization problem for l_1 regularized logistic regression is defined by

$$\min_x \frac{1}{N} \sum_{i=1}^{N} \log(1 + \exp(-b_i x^T a_i)) + \lambda \|x\|_1, \tag{64}$$

where the regularization term $\|x\|_1$ promotes sparse structures on solutions, and where $\lambda > 0$ is the regularization parameter balancing sparsity and fitting error. Problem (64) is a special case of (1) with $f(x) := \frac{1}{N}\sum_{i=1}^N \log(1+\exp(-b_i x^T a_i))$ and $g(x) := \lambda\|x\|_1$. In all the experiments, the matrix B_k in our proximal Newton-type Algorithm 1 is chosen as the Hessian matrix of f at the iteration x^k, i.e., $B_k := \nabla^2 f(x^k)$. We set $\nu := 0.45$ and $\varrho := 2$ in the inexact conditions (14) for determining an inexact solution $\hat{x}^k$ to subproblem (3). We also set $\theta := 0.25$, $\sigma := 0.95$, $\gamma := 0.25$, $C := 2F(x^0)$, $\rho := 2$ in Algorithm 1. As shown in [37, Theorem 8], the subgradient mapping $\nabla f(x) + \partial g(x)$ is metrically subregular at $(\bar{x}, 0)$ for any $\bar{x} \in \mathcal{X}$. Then it can be easily verified that all the assumptions required by Theorem 4.1 are satisfied, and hence the sequence generated by the proposed algorithm for the tested problem (64) locally converges to the prescribed optimal solution with a quadratic convergence rate.

We test here two real datasets "colon-cancer" and "rcv1_train" downloaded from the SVMLib repository[1]. For the colon-cancer dataset, the dimension of the data matrix is $2,000 \times 62$ and is sparse with 124,000 nonzero elements. For the rcv1_train dataset, the dimension of the data matrix is $47,236 \times 20,242$ and is sparse with $1,498,952$ nonzero elements. Both of these two real datasets have more columns than rows, the

[1] http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/.

the loss function f in the corresponding problem (64) is not strongly convex. Since the IRPN proposed in [39] does not require f in (64) to be strongly convex and problem (64) satisfies all the assumptions required by the IRPN, we are going to compare our proposed proximal Newton-type Algorithm 1 with the IRPN. Note that the IRPN code is collected from https://github.com/ZiruiZhou/IRPN. We set $\theta = \rho := 0.25$, $\zeta := 0.4$ and $\eta := 0.5$ in the IRPN as suggested in [39]. Since our interest here is in the local quadratic convergence, we set $\varrho := 1$ in the IRPN. It should be noticed that in such a setting both our Algorithm 1 and the IRPN require solving subproblem (3) with $H_k := \nabla^2 f(x^k) + c\|r(x^k)\|^2$ at each iteration. This subproblem will be solved by the coordinate gradient descent method, which is implemented in MATLAB as a C source MEX-file.[2] We will test the numerical experiments with different values of the parameter c to investigate its impact on the performances of both algorithms.

The initial points in all the experiments are set to be the zero vector. Our Algorithm 1 is terminated at the iteration x^k if the accuracy TOL is reached by $\|r(x^k)\| \leq \text{TOL}$ with the residual $\|r(x^k)\|$ defined via the prox-gradient mapping (8). The maximum number of outer iterations in both algorithms is 50, and the maximum number of iterations for the coordinate gradient descent method to solve the corresponding subproblems is 10000.

The achieved numerical results are presented in Tables 1 and 2. We employ the three values $\{10^{-4}, 5 \times 10^{-4}, 10^{-5}\}$ of the penalty parameter λ and the six values $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$ of the parameter c in the algorithms with reporting the number of outer iterations and the CPU time. Observe from Tables 1 and 2 that our proposed proximal Newton-type Algorithm 1 always achieves the desired accuracy within the preset maximum outer iteration number. The number of iterations and the time taken are almost the same for all the different test values of c. As seen however, the IRPN cannot achieve the required accuracy within the preset maximum outer iteration number for some cases. We plot the corresponding values of the residual $\|r(x^k)\|$ and the step size versus the outer iteration number in Figures 1 and 2 for the some cases where the IRPN fails to achieve the required accuracy. As seen from Figures 1 and 2, when the iteration x^k approaches the optimal solution, the step size of our proposed Algorithm 1 is always 1 and the sequence $\{x^k\}$ exhibit a quadratic convergence rate, while the step size of the IRPN becomes close to 0 and the convergence speed of the sequence $\{x^k\}$ becomes very slow. In fact, when the sequence $\{x^k\}$ generated by the IRPN is close to the optimal solution, the line search strategy in the IRPN rejects the unit step size, and only a small step size is accepted. The poor performance of the IRPN caused by such a small step size can be observed in Figures 1 and 2. In [39, Theorem 1], the authors present a sufficient condition on the value of c to meet a unit step size and hence to guarantee a local quadratic convergence. The validity of this condition relies heavily on the Luo-Tseng error bound radius and the Lipschitz continuity constant of the Hessian $\nabla^2 f$. Unfortunately, the calculation of these parameters of the IRPN are challenging. Theoretically, it is too ambitious to find appropriate values of c. This theoretical limitation agrees with the failures of the IRPN listed in Tables 1 and 2. We can see from both Tables 1 and 2 as well as from Figures 1 and 2 that the proposed Algorithm 1 is stable with respect to the values of c. Furthermore, the line search strategy in Algorithm 1 always accepts a unit step size, and thus the sequence of iterates $\{x^k\}$ generated by Algorithm 1 achieves a fast local convergence rate.

[2] The code is downloaded from https://github.com/ZiruiZhou/IRPN.

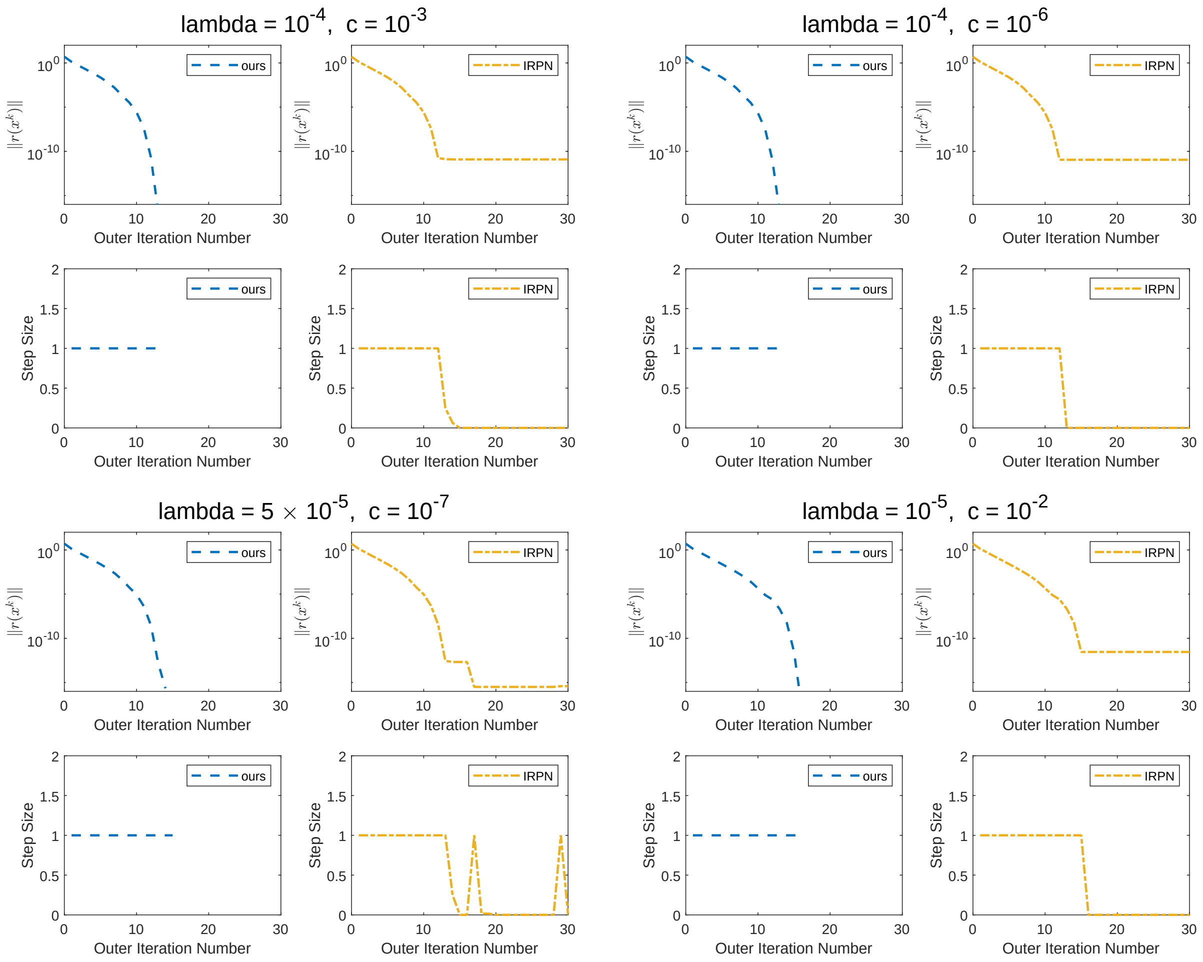

Figure 1: Residual $\|r(x^k)\|$ w.r.t. outer iteration number and step size w.r.t. outer iteration number on colon-cancer dataset

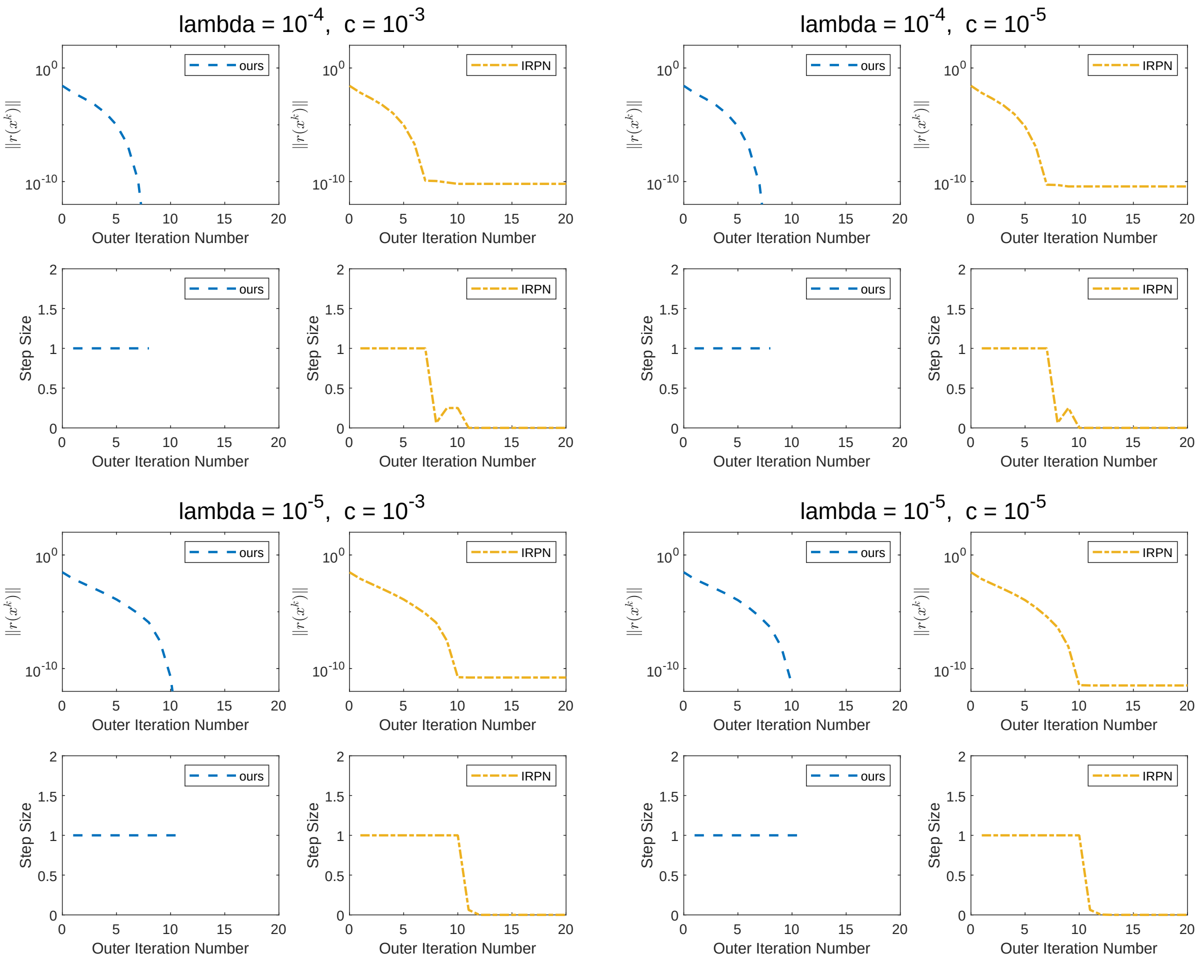

Figure 2: Residual $\|r(x^k)\|$ w.r.t. outer iteration number and step size w.r.t. outer iteration number on rcv1_train dataset

Table 1: Numerical comparison on colon-cancer dataset with TOL $= 10^{-16}$

		$c = 10^{-2}$		$c = 10^{-3}$		$c = 10^{-4}$		$c = 10^{-5}$		$c = 10^{-6}$		$c = 10^{-7}$	
	Solver	ours	IRPN	ours	IRPN	ours	IRPN	ours	IRPN	ours	IRPN	ours	IRPN
$\lambda =$	Outer Iterations	14	$-^{*}$	13	–	13	13	13	–	13	–	13	13
10^{-4}	Time(s)	0.8	–	0.7	–	0.8	0.8	0.8	–	0.8	–	0.8	0.8
$\lambda =$	Outer Iterations	16	15	15	–	14	15	15	15	15	17	15	–
$5 * 10^{-5}$	Time(s)	1.1	1.0	1.0	–	1.0	1.0	1.1	1.1	1.0	6.0	1.0	–
$\lambda =$	Outer Iterations	16	–	16	16	16	16	16	–	16	–	16	–
10^{-5}	Time(s)	1.9	–	1.5	1.5	1.6	1.6	1.4	–	1.5	–	1.5	–

* – indicates the method can not achieve required accuracy TOL with 50 outer iterations.

Table 2: Numerical comparison on rcv1_train dataset with TOL $= 10^{-12}$

		$c = 10^{-2}$		$c = 10^{-3}$		$c = 10^{-4}$		$c = 10^{-5}$		$c = 10^{-6}$		$c = 10^{-7}$	
	Solver	ours	IRPN	ours	IRPN	ours	IRPN	ours	IRPN	ours	IRPN	ours	IRPN
$\lambda =$	Outer Iterations	10	10	8	$-^{*}$	8	–	8	–	8	8	8	–
10^{-4}	Time(s)	8.9	9.0	7.5	–	7.1	–	6.8	–	7.0	7.0	6.8	–
$\lambda =$	Outer Iterations	11	11	9	9	9	9	9	9	9	9	9	9
$5 * 10^{-5}$	Time(s)	14.7	14.6	13.4	13.6	13.2	14.8	13.1	14.4	12.6	16.6	12.9	12.9
$\lambda =$	Outer Iterations	15	–	11	–	11	–	11	–	11	–	11	–
10^{-5}	Time(s)	116.7	–	116.4	–	190.3	–	40.3	–	40.0	–	39.6	–

* – indicates the method can not achieve required accuracy TOL with 50 outer iterations.

References

[1] Aragón Artacho, F.J., Geoffroy, M.H.: Characterization of metric regularity of subdifferentials. J. Convex Anal. 15, 365–380 (2008)

[2] Aragón Artacho, F.J., Geoffroy, M.H.: Metric subregularity of the convex subdifferential in Banach spaces. J. Nonlin. Convex Anal. 15, 35–47 (2015)

[3] Beck, A., Teboulle M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. 2, 83–202 (2009)

[4] Byrd, R.H., Nocedal, J., Oztoprak, F.: An inexact successive quadratic approximation method for L-1 regularized optimization. Math. Program. 157, 375–396 (2016)

[5] Chen, X., Fukushima, M.: Proximal quasi-Newton methods for nondifferentiable convex optimization. Math. Program. 85, 313–334 (1999)

[6] Dan, H., Yamashita, N., Fukushima, M.: Convergence properties of the inexact Levenberg-Marquardt method under local error bound conditions. Optim. Meth. Softw. 17, 605–626 (2002)

[7] Dontchev, A.L., Rockafellar, R.T.: Implicit Functions and Solution Mappings: A View from Variational Analysis, 2nd edition. Springer, New York (2014)

[8] Drusvyatskiy, D., Lewis, A.S.: Error bounds, quadratic growth, and linear convergence of proximal methods. Math. Oper. Res. 43, 919–948 (2018)

[9] Drusvyatskiy, D., Mordukhovich, B.S., Nghia, T.T.A.: Second-order growth, tilt stability, and metric regularity of the subdifferential. J. Convex Anal. 21, 1165–1192 (2014)

[10] Facchinei, F., Pang, J.-S.: Finite-Dimesional Variational Inequalities and Complementarity Problems. Springer, New York (2003)

[11] Fischer, A.: Local behavior of an iterative framework for generalized equations with nonisolated solutions. Math. Program. 94, 91–124 (2002)

[12] Fukushima, M., Mine, H.: A generalized proximal point algorithm for certain non-convex minimization problems. Int. J. Syst. Sci. 12, 989–1000 (1981)

[13] Friedman, J., Hastie, T., Höfling, H. Tibshirani, R.: Psthwise coordinate optimization. Ann. Appl. Stat. 1, 302–332 (2007)

[14] Gaydu, M., Geoffroy, M.H., Jean-Alexis, C.: Metric subregularity of order q and the solving of inclusions. Cent. Eur. J. Math. 9, 147–161 (2011)

[15] Hsieh, C., Dhillon, I.S., Ravikumar, P.K., Sustik, M.A.: Sparse inverse covariance matrix estimation using quadratic approximation. In: Shawe-Taylor, J. et al. (eds) Advances in Neural Information Processing Systems 24, pp. 2330–2338. Curran Associates, New York (2011)

[16] Izmailov, A.F., Solodov, M.V.: Newton-Type Methods for Optimization and Variational Problems, Springer, New York (2014)

[17] Khanh, P.D., Mordukhovich, B.S., Phat, V.T.: A generalized Newton method for subgradient systems. arXiv:2009.10551v1 (2020)

[18] Kruger, A.Y: Error bounds and Hölder metric subregularity. Set-Valued Var. Anal. 23, 705–736 (2015)

[19] Lee, C., Wright, S.J.: Inexact successive quadratic approximation for regularized optimization. Comput. Optim. Appl. 72, 641–0674 (2019)

[20] Lee, J.D., Sun, Y., Saunders, M.A.: Proximal Newton-type methods for minimizing composite functions. SIAM J. Optim. 24, 1420–1443 (2014)

[21] Li, G., Mordukhovich, B.S.: Hölder metric subregularity with applications to proximal point method. SIAM J. Optim. 22, 1655–1684 (2012)

[22] Li, G., Pong, T.K.: Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods. Found. Comput. Math. 18, 1199–1232 (2018)

[23] Luo, Z.-Q., Tseng, P.: On the linear convergence of descent methods for convex essentially smooth minimization. SIAM J. Control Optim. 30, 408–425 (1992)

[24] Mordukhovich, B.S.: Variational Analysis and Applications. Springer, Cham, Switzerland (2018)

[25] Mordukhovich, B.S., Ouyang, W.: Higher-order metric subregularity and its applications. J. Global Optim. 63, 777–795 (2015)

[26] Mordukhovich, B.S., Sarabi, M.E.: Generalized Newton algorithms for tilt-stable minimizers in nonsmooth optimization. arXiv:2004.02345 (2020)

[27] Necoara, I., Nesterov, Yu., Glineur, F.: Linear convergence of first order methods for non-strongly convex optimization. Math. Program. 175, 69–107 (2019)

[28] Nesterov, Yu.: Lectures on Convex Optimization, 2nd edition. Springer, Cham, Switzerland (2018)

[29] Oztoprak, F., Nocedal, J., Rennie, S., Olsen, P.A.: Newton-like methods for sparse inverse covariance estimation. In: Pereira, F. et al. (eds) Advances in Neural Information Processing Systems 25, pp. 755–763. Curran Associates, New York (2012)

[30] Robinson, S.M.: Generalized equations and their solutions, Part II: Applications to nonlinear programming. Math. Program. Stud. 19, 200–221 (1982)

[31] Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control Optim. 14, 877–898 (1976)

[32] Rockafellar, R.T., Wets, R.J-B.: Variational Analysis. Springer, Berlin (1998)

[33] Scheinberg, K., Tang, X.: Practical inexact proximal quasi-Newton method with global complexity analysis. Math. Program. 160, 495–529 (2016)

[34] Schmidt, M., Roux, N., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: Shawe-Taylor, J. et al. (eds) Advances in Neural Information Processing Systems 24, pp. 1458–1466. Curran Associates, New York (2011)

[35] Sra, S., Nowozin, S., Wright, S.J. (eds): Optimization for Machine Learning. MIT Press, Cambridge MA (2011)

[36] Tseng, P.: Approximation accuracy, gradient methods, and error bound for structured convex optimization. Math. Program. 125, 263–295 (2010)

[37] Ye, J.J., Yuan, X., Zeng, S., Zhang, J.: Variational analysis perspective on linear convergence of some first order methods for nonsmooth convex optimization problems. Optimization Online, http://www.optimization-online.org/DB_HTML/2018/10/6881.html (2018)

[38] Yuan, G.-X., Ho, C.-H., Lin, C.-J.: An improved GLMNET for L1-regularized logistic regression. J. Mach. Learn. Res. 13, 1999–2030 (2012)

[39] Yue, M.-C., Zhou, Z., So, A.M.-C.: A family of inexact SQA methods for non-smooth convex minimization with provable convergence guarantees based on the Luo-Tseng error bound property. Math. Program. 174, 327–358 (2019)

[40] Zheng, X.Y., Ng, K.F.: Hölder stable minimizers, tilt stability sand Hölder metric regularity of subdifferentials. SIAM J. Optim. 120, 186–201 (2015)